University “Alexandru Ioan Cuza” of Iași, Romania
Faculty of Computer Science

Natural Language Processing
Using Semantic Frames

Ph.D Candidate
Diana Trandabăț

Scientific Advisor
Prof. Dan Cristea, Ph.D

March 2010
Acknowledgements

Foremost, I would like to thank my advisor, Prof. Dan Cristea, for introducing me to the algorithms and data structures and for accepting the challenge to have a converted Linguist as one of his PhD Students, thus providing me with a unique opportunity to complete a PhD thesis within the Faculty of Computer Science. I am really grateful to him for guiding my steps through the Natural Language Processing field, yet providing me with the freedom to experiment and learn from mistakes.

I also want to thank my (present or past) colleagues from the Natural Language Processing Group within the Faculty of Computer Science, each of them offering me memorable experiences: Adi Iftene, Ionuț Pistol, Maria Husarciu, Alex Moruz, Marius Răschip, Lucian Gadioi, Iustin Dornescu and Corina Forăscu. In larger or smaller, closer or more distant teams, we experienced together different fields of language processing, from e-learning to question answering, sharing nights of last minute rapports, conferences, or cafeteria lines. Very special thanks to my last year’s “office mate”, Corina Dima, for creating the perfect work environment, sweetened with honeyed tea, oranges and relaxing chats, and for accepting mine and my “vicious creature’s” never-ending warbles.

A great thank goes to my colleagues from the Institute of Computer Science within the Romanian Academy, especially to Mr. Curteanu, for treating me as family, for teaching me how to read a scientific paper, and for systemically ordering the research areas to allow me to have the best conditions I could to focus both on my tasks at the Institute, and my PhD thesis.
Many thanks to the colleagues from the Research Institute for Artificial Intelligence within the Romanian Academy (RACAI), especially to Prof. Dan Tufiş, the director of the Institute, and to Radu Ion, Alin Ceasu and Dan Ţtefănescu, who helped me in the development of the semantic roles import method, by aligning the English FrameNet sentences with their Romanian translations. RACAI’s Linguistic Web Services were also very valuable for adding syntactic information to the Romanian sentences, for the semantic role labeling task.

I also acknowledge the financial support offered by two grants of the National University Research Council (CNCSIS) within the Ministry of Education and Research, a scholarship and a research grant for young PhD Students, which provided me with the opportunity to participate at different international conferences and discuss my approaches with known scientists, but also to be introduced in the project management world.

No research could be done without the friends who take you out to celebrate a ring (and the second one), another year gone by, or just a cloudy day. I also thank my friends who kept sending good luck messages and not giving up on me as someone who never responds to emails. To a funny friend, who missed a not-so-regrettable-after-all trip to Greece because of my ever extending PhD, thank you for making me smile when I was covered with to-do’s.

As always, on the last place on paper, but on the first in my heart, I would like to thank my family, for their continuous support. I hope I am making you proud of me, at least as much as you make me proud of you.

The person who should have been the first to mention, since he deserves all my gratitude and love, is my husband, a motivating professional model, and an endless source of encouragements, understanding and ingenuousness to find things to do while keeping me company when working late.
# Contents

List of Figures v

List of Tables vii

Glossary ix

1 Introduction 1

1.1 Semantics in language ................................. 1
1.2 Challenges ............................................. 2
1.3 Thesis structure ........................................ 4

2 What are Semantic Roles? 5

2.1 Linguistic Background ................................. 5
  2.1.1 Frame Semantics .................................. 9
  2.1.2 The Concepts of Prototype and Perspective ........ 11
  2.1.3 Beyond Frame Semantics ............................ 12
  2.1.4 Predicationality .................................... 15
  2.1.5 Valence ............................................. 19
  2.1.6 Case Grammar ...................................... 20
  2.1.7 Semantic Role Theories ............................. 24
  2.2 Description of Semantic Roles ....................... 26
    2.2.1 Common List of Semantic Roles .................. 27
    2.2.2 Characteristics of Semantic Roles ............... 28
    2.2.3 Types of Semantic Roles ........................ 30
  2.3 Semantic Role Resources ............................. 31
    2.3.1 VerbNet ......................................... 31
CONTENTS

2.3.2 FrameNet ................................................. 32
2.3.3 PropBank ................................................. 33
2.3.4 FrameNet vs. PropBank ............................... 36
2.3.5 Combining Semantic Frames with other Lexical Resources 37
2.3.6 Creating Semantic Role Resources for Languages other than
    English ................................................. 41
    2.3.6.1 German ........................................ 42
    2.3.6.2 Japanese ...................................... 43
    2.3.6.3 Spanish ....................................... 44
    2.3.6.4 Adding Romanian to the Semantic Role Resources
        Map .............................................. 44
2.4 Conclusions ................................................. 44

3 Approaches to Semantic Role Labeling 47

3.1 Characteristics of Semantic Role Labelers ............... 49
    3.1.1 Types of Labeled Categories ....................... 49
    3.1.2 Types of Vector Spaces .......................... 50
    3.1.3 Semantic Role Labelers Architecture ............ 53
    3.1.4 Set of Features .................................. 55
3.2 Probability estimation ..................................... 56
    3.2.1 Feature set ...................................... 57
    3.2.2 Estimating probabilities ......................... 60
    3.2.3 Combining probability distributions .............. 61
3.3 Decision trees ............................................. 63
    3.3.1 Feature set ...................................... 63
    3.3.2 Feature Salience ................................ 66
3.4 Support Vector Machines (SVMs) approaches ............ 67
    3.4.1 Feature Set ...................................... 69
    3.4.2 Feature Salience ................................ 71
3.5 Memory Based Learning (MBL) ............................. 72
3.6 Conclusions ................................................. 75
## 4 Development of a Semantic Role Labeling System

4.1 RuleSRL: A Rule-based Semantic Role Labeling System ............................. 78
  4.1.1 Input data ................................................................................. 78
  4.1.2 The System’s Architecture ...................................................... 83
    4.1.2.1 Predicate Identification ...................................................... 83
    4.1.2.2 Predicate Sense Identification ............................................ 86
    4.1.2.3 Semantic Roles Identification .......................................... 92
  4.1.3 Results .................................................................................... 94

4.2 PASRL: a Platform for Adjustable Semantic Role Labeling ......................... 97
  4.2.1 Weka Classifiers .................................................................... 97
  4.2.2 Training data ......................................................................... 100
  4.2.3 Learning Steps ...................................................................... 100
    4.2.3.1 Predicate Prediction .......................................................... 102
    4.2.3.2 Joint Predicate and Predicate Sense Identification ............ 106
    4.2.3.3 Argument Prediction .......................................................... 107
    4.2.3.4 Considerations on data sets ................................................ 110
  4.2.4 Results .................................................................................... 111
  4.2.5 Using PASRL pre-trained models .............................................. 114
  4.2.6 Using PASRL for Romanian .................................................... 114

4.3 Conclusions .................................................................................. 115

## 5 Romanian Resource of Semantic Roles

5.1 Intuition ...................................................................................... 117

5.2 System Architecture .................................................................... 121
  5.2.1 The Translation Module ........................................................... 123
  5.2.2 Alignment Module .................................................................. 125
  5.2.3 Import Module ...................................................................... 128
    5.2.3.1 One-to-zero import ............................................................. 131
    5.2.3.2 One-to-one import ............................................................. 131
    5.2.3.3 One-to-many import ............................................................. 131
    5.2.3.4 Many-to-one import ............................................................. 132
    5.2.3.5 Zero-to-one import ............................................................. 132
  5.2.4 Validation Module .................................................................. 133
CONTENTS

5.3 Results Analysis ........................................... 133
   5.3.1 Double Annotation ................................. 134
   5.3.2 Imbrications ........................................ 135
   5.3.3 Unexpressed Semantic Frames .................. 136
5.4 Conclusions .............................................. 137

6 Applications of Semantic Role Labeling Systems 139
   6.1 Semantic Roles in Question Answering ............ 139
   6.2 Semantic Frames for Prosody Generation ............ 144
   6.3 Conclusions ...................................... 151

7 Conclusions and Further Work ............................ 153

References ............................................... 159

Appendices ............................................... 177
List of Figures

2.1 Commercial Transaction frame ......................................... 7
2.2 Redefined Predication ..................................................... 18
2.3 PropBank ................................................................. 35
3.1 Example of Parse Tree ..................................................... 57
3.2 Path Feature ............................................................... 59
3.3 Feature distribution lattice ............................................... 62
3.4 Content word vs. head word ............................................. 65
4.1 Example of Dependency Relations ....................................... 81
4.2 Example of Dependency Tree ............................................ 81
4.3 Rule-based SRL System Architecture .................................. 84
4.4 Example of PropBank Annotation ........................................ 87
4.5 Example of Semantic Role Annotation .................................... 94
4.6 Evaluation of the rule-based SRL System for the Predicate Identification / Predicate Sense Identification module ................................................. 95
4.7 Evaluation of the rule-based SRL System for the Argument Identification task ................................................................. 96
4.8 Architecture of the PASRL Platform ...................................... 101
5.1 Architecture of the system ................................................ 122
5.2 FrameNet Semantic Role Annotation Format .......................... 123
5.3 Translation Interface of the Semantic Role Annotation Import Program ................................................................. 125
6.1 Architecture of the UAIC QA system .................................... 140
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Example of probabilities computed by Gildea and Jurafsky (2002b) for the training set. The governing category feature is defined only for noun phrases.</td>
<td>61</td>
</tr>
<tr>
<td>4.1</td>
<td>Rule-based SRL system pre-processed input format</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>Example of Syntactic and Dependency parsed sentence</td>
<td>82</td>
</tr>
<tr>
<td>4.3</td>
<td>ML algorithms evaluated using 10-fold cross-validation for English, for the Predicate Identification task</td>
<td>104</td>
</tr>
<tr>
<td>4.4</td>
<td>ML algorithms evaluated using 10-fold cross-validation for English, for the Predicate Sense Identification task</td>
<td>106</td>
</tr>
<tr>
<td>4.5</td>
<td>ML algorithms evaluated using 10-fold cross-validation for English, for the Sense Identification task</td>
<td>107</td>
</tr>
<tr>
<td>4.6</td>
<td>ML algorithms evaluated using 10-fold cross-validation for English, for the Argument Identification task</td>
<td>109</td>
</tr>
<tr>
<td>4.7</td>
<td>Best 5 ML algorithms evaluated using 10-fold cross-validation for English, considering the whole training set, for the Predicate Identification task</td>
<td>112</td>
</tr>
<tr>
<td>4.8</td>
<td>Best 5 ML algorithms evaluated using 10-fold cross-validation for German, considering the whole training set, for the Predicate Identification task</td>
<td>112</td>
</tr>
<tr>
<td>4.9</td>
<td>Best 5 ML algorithms evaluated using 10-fold cross-validation for Czech, considering the whole training set, for the Predicate Identification task</td>
<td>112</td>
</tr>
</tbody>
</table>
LIST OF TABLES

4.10 Best 5 ML algorithms evaluated using 10-fold cross-validation for Chinese, considering the whole training set, for the Predicate Identification task ................................................. 113

4.11 Best 5 ML algorithms evaluated using 10-fold cross-validation for Japanese, considering the whole training set, for the Predicate Identification task ................................................. 113

5.1 Example of Word-level Alignment using the aligner developed at RACAI ............................................................... 127
Glossary

Terminology

case grammar Linguistic theory, introduced by Ch. Fillmore in 1968, that analyses the surface syntactic structure of sentences by focusing on the link between the valence of a verb and the semantic context it requires.

frame element Frame-specific semantic role, annotated in the FrameNet corpus.

frame semantics A descriptive framework for characterizing lexical meaning in terms of semantic frames.

FrameNet A lexicographic research project, that produced a lexicon containing very detailed information about the syntax semantics relations of the English predicational words (verbs, nouns and adjectives).

lemma A unit composed of one or more lexemes seen as bearing one or more senses, e.g. the lemma bring up consists of the lexemes bring and up.

lexeme A word in a given part of speech, instantiated by one or more word-forms, e.g. the lexeme bring has the word forms bring, brings, bringing, brought.

lexical unit A word or a sense of a polysemantic word, for which semantic proprieties are defined.

predicational word A word (usually a verb, but also a noun or an adjective) is called predicational if its meaning involves a process event or process name.

PropBank The Proposition Bank project adds a layer of semantic roles to the syntactic structures of the Penn Treebank resource.

prosody The “melody of the speech”, the property that makes natural voice pleasant, coherent, non-metalic.

semantic annotation The process of assigning semantic role tags to the syntactic constituents of a sentence.
**semantic frame**  A semantic argument-predicate structure, linked by linguistic conventions to the meaning of lexical units.

**semantic role** (also semantic case, thematic role, or deep case)  Each element of a semantic frame, corresponding to a syntactic constituent.

**semantic role resource**  A collection of sentences annotated with semantic roles.

**target word**  The predicational word under consideration, and in respect to which annotation is provided.

**WordNet**  a large freely and publicly available lexical database of English, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG</td>
<td>Construction Grammar</td>
</tr>
<tr>
<td>FCG</td>
<td>Fluid Construction Grammar</td>
</tr>
<tr>
<td>FE</td>
<td>Frame Elements, the notation of Semantic Roles in FrameNet</td>
</tr>
<tr>
<td>FN</td>
<td>FrameNet</td>
</tr>
<tr>
<td>HPSG</td>
<td>Head-driven Phrase Structure Grammar</td>
</tr>
<tr>
<td>IE</td>
<td>Information Extraction</td>
</tr>
<tr>
<td>LU</td>
<td>Lexical Unit</td>
</tr>
<tr>
<td>NE</td>
<td>Named Entity</td>
</tr>
<tr>
<td>NG or NP</td>
<td>Noun Group or Noun Phrase</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>OBJ</td>
<td>Syntactic Object of an verb</td>
</tr>
<tr>
<td>OBJD</td>
<td>Syntactic Direct Object of an verb</td>
</tr>
<tr>
<td>OBJI</td>
<td>Syntactic Indirect Object of an verb</td>
</tr>
<tr>
<td>PB</td>
<td>PropBank</td>
</tr>
<tr>
<td>POS</td>
<td>Part of Speech of a word</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>PP</td>
<td>Prepositional Group</td>
</tr>
<tr>
<td>PREDF</td>
<td>Predicational feature of a word (verb, noun or adjective)</td>
</tr>
<tr>
<td>QA</td>
<td>Question Answering</td>
</tr>
<tr>
<td>SO</td>
<td>Systemic Order</td>
</tr>
<tr>
<td>SR</td>
<td>Semantic Role</td>
</tr>
<tr>
<td>SRL</td>
<td>Semantic Role Labeling</td>
</tr>
<tr>
<td>SUBJ</td>
<td>Syntactic subject of a verb</td>
</tr>
<tr>
<td>SVO</td>
<td>Subject-Verb-Object</td>
</tr>
<tr>
<td>TFA</td>
<td>Topic - Focus Algorithm</td>
</tr>
<tr>
<td>VG or VP</td>
<td>Verbal Group or Verbal Phrase</td>
</tr>
<tr>
<td>WN</td>
<td>WordNet</td>
</tr>
<tr>
<td>WSD</td>
<td>Word Sense Disambiguation</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language, widely used as a format for the exchange of data between different computer systems, programs, etc.</td>
</tr>
</tbody>
</table>
Introduction

1.1 Semantics in language

Linguists have always attempted at reconstructing the organization of natural languages in terms of structural notions. Saussure (1916) introduced the dichotomy of the linguistic sign, as being formed through an indissoluble link between a *signifiant*, or phonetic signifier, and a *signifié*, or signified concept. A topical division in linguistics distinguishes therefore between the study of the language structure (Syntax) and the study of the language meaning (Semantics). In the following years, Pragmatics appeared as a reaction to Saussure’s structuralist linguistics, expanding upon his idea that language has an analyzable structure, composed of parts that can be defined in relation to others. These three basic layers of grammatical organization are summarized as:

**Syntactic relations** (e.g. Subject or Object), which define different perspectives through which states of affairs are presented in linguistic expressions;

**Semantic relations** (e.g. Agent, Goal, Recipient, etc.), which define the roles that participants play in states of affairs, as designated by predications;

**Pragmatic relations** (e.g. Topic or Focus, Theme or Rheme), which define the informational status of linguistic expressions, as used in given contexts.

The main domain this thesis covers is Natural Language Semantics, whose basic area of examination is the meaning of signs, and the study of the relations
between different linguistic units, such as: homonymy, synonymy, antonymy, polysemy, hypernymy, hyponymy, meronymy, metonymy, holonymy, linguistic compounds, etc. A key concern in this domain is the identification of the mechanism that allows the attachment of meaning to larger chunks of text, possibly as a result of the composition from smaller units of meaning. Natural Language Semantics includes therefore the study of sense and denotative references, argument structures, semantic roles, discourse analysis, and the linking of all of these to syntax.

Semantic role analysis intends to answer the following questions: How do entities carry out events? What roles do entities play in different events? A semantic role represents the relationship between a predicate and an argument. In any language, verbs can be grouped together into semantic classes, which share common elements of meaning. It is believed that verbs in such classes will also share some syntactic characteristics, since, according to Levin (1993), the semantics of a verb determines, at least partially, its syntactic behavior.

Semantic parsing, by identifying and classifying the semantic entities in context and the relations between them, has great potential on its downstream applications, such as text summarization, question answering, and machine translation. As a result, semantic parsing can be an important intermediate step for natural language understanding.

1.2 Challenges

The main question this thesis intends to answer to is if semantic role information is cross-linguistically valid, and if so, up to what extent. The interest begun when observing the huge amount of time and human resources involved in creating the semantic role resources for English, and later for German, Spanish and Japanese. Since semantic information is considered of major influence for a natural language processing system, we started to consider developing such a resource for Romanian, but with considerable less human and temporal resources. We therefore investigated the transfer of semantic role annotation from English to Romanian. After this resource is built, an automatic role labeler can be created to annotate raw text with the semantic role information, for usage in natural language
1.2 Challenges

processing tools.

The challenges this thesis is attempting to cover are:

**Introducing** a method to create a semantic role annotated corpus with minimum resources (more than 50 persons were involved in the development of the English FrameNet, and the resource is being developed and improved since 1997). The method is applied to Romanian, and we believe it can be successfully used for other language;

**Investigating** the cross-linguistic properties of semantic roles;

**Developing** a semantic role labeling system for English and Romanian;

**Improving** natural language processing tools by using semantic role information;

The major contributions of this thesis to the natural language processing field can be summarized as:

- Establishing a semantic role import method that uses a resource created for one language and transfers the semantic role annotation to another language;

- Creating a resource of annotated Semantic Roles for Romanian by using this method;

- Providing an interface that can be used to easily further develop the semantic role resource;

- Creating a platform for developing adjustable supervised semantic role labelers. This platform trains different machine learning algorithms on a training set, selects the best and builds a semantic role labeler using the best models;

- Testing the platform by creating a Semantic Role Labeler for English and Romanian;

- Contributing to NLP applications by integrating semantic frames in a Question-Answering system and by designing the improvement strategy of a Prosody generation system based on semantic roles.
1. INTRODUCTION

1.3 Thesis structure

This thesis is structured into seven chapters: Chapter 2 presents an overview of the theories on semantic frames, giving the definition and characteristics of semantic roles. Linguistically interpreted corpora are the starting point of supervised machine learning paradigms of natural language processing. The information encoded in the corpora determines to a large extent what can be learned by supervised machine learning systems. Therefore, it is crucial to encode the desired level of information for its automatic acquisition. Since Semantic Roles provide a general semantic interpretation of the sentence, researchers have started in the last decade to build resources capturing the behavior of Semantic Roles. The most important ones are presented at the end of Chapter 2.

Chapter 3 presents several techniques used in order to automate the process of semantic role acquisition. The chapter starts with the presentation of the “classical” architecture of semantic role labelers, mentions then a list of the most used features, ending with the presentation of several automatic role labelers that achieve state-of-the-art performance.

Chapter 4 describes a semantic role labeling system development for English and other languages that have developed semantic role resources. The system’s performance was evaluated on the ConLL2009 shared task corpus.

Chapter 5 presents the development of a resource of semantic roles for Romanian. The starting point for the German, Japanese and Spanish semantic role resource creation was the manual annotation of corpora existing in each language. However, this is a very time and resource consuming task, so for Romanian, we propose creating a corpus of semantic roles starting from the translation of (a part of) the English corpus of annotated sentences.

Chapter 6 discusses the application of automatic annotation of semantic roles in natural language processing systems. The envisaged applications are question answering and prosody generation.

The thesis concludes with a Discussion section, including some further considered developments and an emphasis on the contributions of this thesis to the field of Semantic Role Labeling.
What are Semantic Roles?

2.1 Linguistic Background

All content elements of a language are seen as predicates, i.e. expressions which designate events, properties of, or relations between, entities. The predication represents the mechanism that allows entities to instantiate properties, actions, attributes and states. Linguistic expressions can be dependent or independent. The dependent linguistic expressions are usually different phenomena, while the independent ones are individuals. For example, the word hat can be understood outside any circumstance, time, or person, because it does not have to be attributed to anything or anyone: it is independent, thus an individual. On the contrary, if we consider the word red, the denotations for this word cannot be understood outside its association with an individual: red hat. In linguistic terms, the dependent phenomena are predicates, while individuals are arguments. The linking between a phenomenon and individuals is known as predication.

Predicates are not treated as isolated elements, but as structures, named predicate frames (Dik, 1987) or semantic frames (Chomsky, 1965). Within the predicate frames, each entity (frame element) plays a role, called thematic role (Frawley, 1992), semantic case (Fillmore, 1968), semantic role (Dillon, 1977)\(^1\), thematic relation (Gruber, 1965; Jackendoff, 1990) or, from a purely syntactic perspective, \(\theta\)-role Chomsky (1965). Semantic roles represent in fact the seman-

\(^1\)Semantic role is the name we will use throughout this thesis to refer to the roles entities play within a sentence.
2. WHAT ARE SEMANTIC ROLES?

tic relations that connect individuals to phenomena, or in the linguistic terms, arguments to predicates. After establishing the semantic relations within the predicate frames, syntactic and pragmatic functions are added to each predicate frame element.

Semantic relations are one of the oldest classes of constructors in linguistic theory, the first attempt of providing a theory of linking morphology and semantics being considered to date back thousands of years to Panini’s kāraka theory (Misra, 1966). Longevity, in this case, begets variety, and the literature records various proposals for different sets of semantic roles. These sets of roles range from very specific to very general, and many have been used in computational implementations of one type or another.

The semantic relations can be exemplified within the Commercial Transaction Frame, whose actors include a buyer, a seller, goods, and money. Among the large set of semantically related predicates, linked to this frame, we can mention buy, sell, pay, spend, cost, and charge, each of which indexes or evokes different aspects of the frame. The verb buy focuses on the buyer and the goods, backgrounding the seller and the money; sell focuses on the seller and the goods, backgrounding the buyer and the money; pay focuses on the buyer, the money, and the seller; backgrounding the goods; and so on. The idea is that knowing the meaning of any of these verbs requires knowing what takes place in a commercial transaction and, to some extent, knowing the meaning of all the predicates involved in the frame. The knowledge and experience structured by the Commercial Transaction frame provide the background and motivation for the categories represented by these verbs. The annotation of the Commercial Transaction frame is presented in figure 2.1.

A complete description of the predicates in the commerce frame may also include information about their grammatical properties and the various syntactic patterns in which they occur (their subcategorization frames\(^2\)). Which frame elements may be realized as the subject of the verb, or as its object, if there is one, and what will be the syntactic surface form of the other frame elements? Which ones of these elements are optional and which are mandatory? For example, in

\(^2\)A definition of the subcategorization feature is provided in (Cornilescu, 2006): “A subcategorization feature indicates the (minimal) frame in which some lexical item is allowed to be inserted.”
2.1 Linguistic Background

Commercial_transaction

Definition:
These are words that describe basic commercial transactions involving a Buyer and a Seller who exchange Money and Goods. The individual words vary in the frame element realization patterns. For example, the typical patterns for the verbs buy and sell are: Buyer buys Goods from the Seller for Money. Seller sells Goods to the Buyer for Money.

FEs:

Core:
- **Buyer (Byr)**: The Buyer wants the Goods and offers Money to a Seller in exchange for them.
- **Goods (Gds)**: The FE Goods is anything (including labor or time, for example) which is exchanged for Money in a transaction.
- **Money (Mny)**: Money is the thing given in exchange for Goods in a transaction.
- **Seller (Sllr)**: The Seller has possession of the Goods and exchanges them for Money from a Buyer.

Non-Core:
- **Means (Mens)**: The means by which a commercial transaction occurs.
  - Abby **BOUGHT** the car **with cash**.
  - Robin **PAID** for the car **by check**.
- **Rate (Rate)**: Price or payment per unit of Goods.
  - The manager **PAY**s the paper boys five dollars an hour **Rate**.
- **Unit (Un)**: The Unit of measure of the Goods according to which the exchange value of the Goods (or services) is set. Generally, it occurs in a by-PP.
  - Bob **SELLS** peppers **by weight**.
  - Lee **BUYS** potatoes **by the pound**.
  - She is **PAID** **by the hour**.

Figure 2.1: Commercial Transaction frame - Example of semantic frame with annotated Semantic Roles, extracted from the FrameNet resource
2. WHAT ARE SEMANTIC ROLES?

the sentence:

(1) Carla bought the computer from Sally for $100.

there are four frame elements for the predicate buy: Carla, the computer, from Sally and for $100. The subject, Carla, is the buyer and the direct object, the computer, represents the goods; both elements being mandatory for the predicate meaning to be completed. If any of these mandatory roles were missing from the sentence, the informational content of the sentence would be incomplete and the message could not be successfully transmitted without further clarifications (imagine someone comes into a room and announces only bought the computer or Carla bought). The other two frame elements are optional, and surface as oblique objects: from Sally is the seller, and for $100 represents the money (the price). Their lexicalization in the sentence is only an complementary information, and neither the presence or absence of this information affects the primary message concerning Carla’s acquisition of a computer. An important observation is that different prepositions are used with semantic role. For example, from is the preposition allowing the interpretation of Sally as the seller, while for the money role, either for or with can be used. The syntactic-semantic description of each predicate forms the predicate subcategorization frame. This information is not mandatory for the semantic frame of a predicate, but it is nevertheless deductible from the descriptions of the different frame elements in large annotated semantic role resources.

Considering a more general money-transferring frame serves as a reminder of the many nouns whose meanings may be described in frame semantic terms - i.e. with respect to the underlying conceptual framework needed for their understanding and with reference to the knowledge and experience the speakers has of the background situations. Among the nouns linked to a money-transferring frame are tip, ransom, allowance, refund, honorarium, bounty, tuition, retainer, bonus, rent, fare, child support, bus money, salary, reward, and alimony. Referring to a sum of money with any one of these nouns requires rich and detailed information about a much larger scene than the Commercial transaction frame, of which the actual transferring of money is just a small part. For example, using the word alimony assumes something like the following: two people who used to be married are now divorced; upon divorce it is agreed that
one of them gives the other a sum of money at regular intervals, usually monthly. There is, of course, much more involved, including legal negotiations, court decrees, and so on. The claim being made here is that each predicate (either verb or noun) bring along an entire scene.

Among other factors, considered for a full description of nouns linked to a money-transferring frame, is the syntax of the expressions in which a particular noun occurs. Some nouns require the indefinite article, while others require a possessive pronoun. The choice depends on whether the money transferred is expected and whether talk about that money takes place before or after an agreement about the transfer. If the money is not expected, the indefinite article can be used: She gave him a tip/a reward/a bonus. If the money is expected, the possessive pronoun is needed: She gave him his allowance/his salary/his change. In order to capture, for each frame, the different syntactic and semantic combinations, linguists have begun to annotate large corpora with syntactic/semantic information. These resources, presented in Section 2.3, have been then used to train different machine learning programs in order to develop automatic annotation software which enrich the texts with semantic information.

2.1.1 Frame Semantics

The use of predicate frames and their semantic relations has conveyed into the Frame Semantics theory. If in the linguistic theories existent at that moment, generically called by Fillmore (1975) “checklist theories of meaning”, the meaning of a linguistic form is represented in terms of a checklist of conditions that have to be satisfied in order for the form to be appropriately or truthfully used, in Frame Semantics (Fillmore, 1985), word meaning is characterized in terms of experience-based schematizations of the speaker’s world - i.e. frames which impose order on frame elements. Frame Semantics also contrasts with the theories on semantic fields as characterized and practiced by field theorists. In semantic field theory, a word is defined in terms of relation to its peers, i.e. other words in the same field. In Frame Semantics, a word is defined in relation to its background frame, not in relation to other words. A word’s meaning depends on its conceptual environment, knowledge of which is necessary for its appropriate use.

Much of the Frame Semantics literature covers frames and individual words
2. WHAT ARE SEMANTIC ROLES?

(or sets of words) and expressions. In addition to its utility in lexical semantics, the frame is also considered a useful tool in text semantics and the semantics of grammar. The interpreter of a text invokes a frame when assigning an interpretation to a piece of text by placing its contents in a pattern known independently of the text. A text evokes a frame when a linguistic form or pattern is conventionally associated with that particular frame. For example, consider the sentence:

(2) Julia will open her presents after blowing out the candles and eating some cake.

Although there is no mention of a birthday party, interpreters sharing the requisite cultural background invoke a birthday party scene. With the noun phrase birthday presents, instead of just presents, the sentence contains words which evoke the very same scene. Using the frame concept in text semantics establishes and affirms the close link between lexical semantics and text semantics not only because a lexical item can be taken as a very small text, but also because the meaning of any single lexical item plays an important role in the construction of the meaning of any (longer) text, for example, a sentence. To illustrate, consider the preposition: on, as in:

(3) The children played on the bus.

The sentence (3) describes a scene in which some children were playing while the bus was in operation as a vehicle for transportation. That sentence could not be used appropriately to describe a scene in which the children were playing in an abandoned wheelless bus in a vacant lot, for which only the sentence:

(4) The children played in the bus.

is appropriate. The construction of the meaning of the first sentence relies on more than just understanding the basic meaning of the preposition on. It relies on knowledge of the details of the situation framed by on, particularly that the bus has to be “in service...and...destined to make the journey anyway” (Fillmore, 1985). Such an approach to text semantics highlights the contribution of word meaning, defined in terms of frames, to sentence interpretation.
2.1.2 The Concepts of Prototype and Perspective

A number of important concepts figure into the Frame Semantics (Fillmore, 1985) approach to linguistic description and analysis. One such concept is that of a prototype (Bornkessel et al., 2006), understood as a fairly large slice of the surrounding culture against which the meaning of a word is defined and understood. For example, to understand the meaning of the word breakfast, it is necessary to understand the institutions and practices of the culture in which this category exists. In this case, it is necessary to understand the practice of eating three meals a day at more or less fixed times and that the meal eaten in the early part of the day after a period of sleep has a special menu; for this meal, the word breakfast is used. The conditions which define the prototype need not all be present in order for native speakers to use the word appropriately. Speakers of American English may use the word breakfast for the meal eaten in each of the following situations: sleeping through the morning, eating eggs, toast, coffee, and orange juice at two in the afternoon; staying up all night, eating eggs, toast, etc. at seven in the morning; sleeping through the night, eating a peanut butter and jelly sandwich at seven in the morning. This range of usage can be captured in an account of word meaning which appeals to the notion of a prototype. The word breakfast provides a category which can be used in a variety of contexts; the contexts are determined by the word’s prototypic use.

Defining words in terms of frames and prototypes provides a useful approach to the boundary problem for linguistic categories. The word bachelor, an often-cited example in the literature, is defined against a prototype background frame rather than in terms of all the unusual circumstances in which the word might be used. That bachelor might occur in contexts which don’t match the prototype suggests that speakers are willing to extend the word’s frame or create a new frame.

Another important concept in Frame Semantics is that of perspective. For example, let’s consider the sentence:

(5) Carla bought the computer from Sally for $100.

which evokes the Commercial Transaction frame discussed above. This sentence takes the perspective of the buyer. Similarly, the sentence:
2. WHAT ARE SEMANTIC ROLES?

(6) Sally sold the computer to Carla for $100.

is a report of a commercial event from the perspective of the seller. That the notion of perspective depends, to a certain extent, on world knowledge, can be illustrated with the words *land* and *ground*, both of which identify the same entity, the dry surface of the earth. Hearing that a traveler spent a few hours on *land*, we understand that the traveler interrupted a *sea voyage*; hearing that a traveler spent a few hours on *the ground*, we understand that the traveler interrupted an *air flight*. Thus, the different words assume different perspectives on or schematizations of the same scene; understanding the choice of words for talking about that scene requires appealing to the history of events leading up to it.

2.1.3 Beyond Frame Semantics

Frame Semantics takes as a goal a uniform representation for the meanings of words, sentences, and texts. Indeed the labeled box notation initially suggested as an informal representation system for the lexicon was refined and used for the representation of grammatical constructions in the grammatical framework developed by Fillmore and his colleagues, Construction Grammar (Kay and Fillmore, 1999). The connection between Frame Semantics and Construction Grammar goes beyond the matter of representation. Construction Grammar views the description of grammatical patterns and the semantic and pragmatic purposes they serve as equally important and necessary. In Construction Grammar, the semantic frame associated with a lexical item provides some of the semantic information needed for the semantic interpretation of a sentence. As with lexical items and texts, semantic descriptions and explanations of grammatical constructions appeal to frames for background information about the scene organized or schematized by the construction.

Alternative approaches to formalize construction grammar using syntax-to-semantic mapping led to the Fluid Construction Grammar (FCG), a linguistic formalism designed to explore how far a construction grammar approach can be used for handling open-ended grounded dialogue, i.e. dialogue between or with autonomous embodied agents about the world as experienced through their sensory-motor apparatus. The FCG formalism (Steels and De Beule, 2006) builds
2.1 Linguistic Background

heavily on the tradition of unification-based feature structure grammars such as HPSG (Pollard and Sag, 1994).

FCG grew out of efforts to understand the creative basis of language. Steels believed that “we need to understand how new aspects of language (new concepts and conceptualizations, new lexical items, new syntactic and semantic categories, new grammatical constructions, new interaction patterns) may arise and spread in a population, the same way biologists try to understand how new life forms may arise” (Steels, 2003).

In FCG, the information about an utterance is organized in a semantic and a syntactic structure. The semantic structure is a decomposition of the utterance’s meaning and contains language-specific semantic re-categorizations (for example a put-event is categorized as a cause-move-location with an agent, a patient and a location). The syntactic structure is a decomposition of the form of the utterance into constituents and morphemes and contains additional syntactic categorizations such as syntactic features (like number and gender), word order constraints, etc. FCG organizes the information about an utterance in feature structures, similar to other feature-structure based formalisms.

There is a strong correspondence between the syntactic and semantic structure built up for the same utterance, although there can be units which only appear in the syntactic structure (for example for grammatical function words) and vice versa. A FCG rule (also called template) typically expresses constraints on possible meaning-form mappings. A rule has two poles: a left pole which typically contains constraints on semantic structure formulated as a feature structure with variables, and a right pole which typically contains constraints on syntactic structure again formulated as a feature structure with variables. Rules are divided into rule subsets which help constrain the order of rule-application and design large-scale grammars. Thus, FCG makes a distinction between morph-rules, which decompose a word into a stem and pending morphemes and introduce syntactic categories; lex-stem-rules, which associate meaning with the stem as well as valence information and a role-frame; con-rules, which correspond to grammatical constructions that associate parts of semantic structure with parts of syntactic structure; and sem and syn-rules which perform inference over semantic or syntactic categories to expand semantic or syntactic structure.

Frame semantics has also been used to account for different sorts of syntactic
2. WHAT ARE SEMANTIC ROLES?

Phenomena. For example, Lambrecht (1984) offers an analysis of formulaic German binomial expressions of the form *N und N* employing frames to account for their creation and use, which is subject to semantic and pragmatic constraints. In general, German noun phrases (NPs) headed by countable nouns require determiners (*Das Mädchen und ihr Hund* - 'the girl and her dog'). However, in certain more or less fixed combinations, such as *Messer und Gabel* - 'knife and fork', *Hut und Mantel* - 'hat and coat', as well as in novel pairs such as in *Schule und Arbeitswelt* - 'in school and at work', such NPs may go without determiners. Lambrecht argues that such bare binomial expressions are licensed whenever the paired noun can be construed as belonging to a single semantic frame. This frame may be justified by historical, social, or cultural contexts. German bare binomial constructions demonstrate the relevance of semantic frames for syntactic well-formedness.

In another work involving issues of syntactic well-formedness, Lakoff (1986) argues against a purely syntactic account of the coordinate structure constraint and suggests that the extraction phenomenon can be explained by appealing to frame semantics. More specifically, Lakoff proposes a notion of a “natural course of events” characterized in terms of a semantics of understanding (Fillmore, 1985), rather than truth-conditional semantics. Lakoff’s “natural courses of events” or “scenarios” are “humanly-constructed holistic organizations of states and events” - i.e. instances of frames. The analysis offers a way of using frames to account for a syntactic phenomenon which is determined, in part, by semantic and pragmatic criteria.

Adopting a construction grammar approach to the analysis of argument structure constructions, Goldberg (1995) also uses the frame idea. Basic argument structure types are defined relative to background frames/scenes of highly structured cultural and world knowledge. That is, argument structure constructions (e.g. the ditransitive construction, the caused-motion construction, the resultant construction, etc.) also invoke frames which designate event types fundamental to human experience. Thus, the meaning of a particular argument structure construction derives not only from the meaning of the particular verb in the construction, but also from the construction itself, whose meaning derives, in part, from the frame-semantic knowledge with which it is associated. By way of

---

3 Compound noun structures of the type “noun and noun”.

example, the ditransitive construction involves a scene in which someone gives something to someone. The claim is that the meaning of a sentence such as Sara faxed Jeremy the invoice derives not simply from the meaning of the verb fax, but also from what speakers of a language know about a situation where one person gives something to another person. Among other things, his work demonstrates the usefulness of frames in relating world knowledge and language structure.

2.1.4 Predicationality

Some predicates demand semantic frames around them, while others don’t. This behavior is commanded by a feature called Predicalinality, considered to exist at the lexical level for the major lexical categories noun (N), verbs (V), and adjectives (A), corresponding to what in the literature is usually called the deverbality property, or the deverbality of these categories. For an extended survey and analysis of the notion and its syntactic-semantic consequences, see Curteanu (2003); Curteanu et al. (2006b). The term deverbality is avoided because its meaning is not necessarily specific to verbs, since this essential lexical feature is equally shared by verbs, nouns and adjectives. Moreover, there are classes of verbs which do not bear this property, e.g. the auxiliaries or the support verbs. The feature of Predicalinality is assigned to finite or non-finite verbs, nouns, and adjectives, whose meaning involves a process event or process name. This feature is abbreviated as PREDF (Predicational Feature), with two main values: [+predicational] for words indicating active processes, or [-predicational] for words indicating states. A word is considered to have the predicationality feature if it demands an argument structure in order to complete its sense.

The feature of predicationality, as a lexical semantics quality, is not necessarily related to the predicate (which is a syntactic construction): in the nominal predicate, the copulative verb is not a predicational verb.

(7) This [is]_{PREDF} my house.

The same goes for the auxiliaries and support verbs incorporated within a verbal group, since they act as functional categories, which lack assigning properties, since they do not designate events. Corniles cu (1994) considers that these verbs
2. WHAT ARE SEMANTIC ROLES?

rather specify “the reference of the event expressed by the main verb, as members of the main verb’s Tense Chain”.

(8) a. I \texttt{[shall]}_{-PREDF} \texttt{go}.
    b. I \texttt{[used]}_{-PREDF} to go there.

This does not exclude, in the nominal predicate, that the predicative nominal (as semantic head of the construction) bears the feature of predicationality, as the predicative nominals 'explanation', 'marking', 'receiving' etc. (which are predicational nouns) in the nominal predicates of the clauses.

(9) This \texttt{[is]}_{-PREDF} his \texttt{[explanation]}_{+PREDF} of the situation.

The argument structure around a predicational word generally involves the following syntactic constituents: the subject, the direct and the indirect object. Due to the existence of an agreement connection between the subject and the predicate of a sentence, the classical view over predication involves only the pair Subject - Predicate.\footnote{The subject is what (or whom) the sentence is about, while the predicate tells something about the subject.}

However, when considering the predicational feature and the semantic space, the verbal group \footnote{For the Romanian language, an investigation of the verbal group and its constitutive elements are presented in (Curteanu et al., 2006b), while the lexical predication and its logical representation are discussed in (Curteanu et al., 2006a).} – or verbal complex in Monachesi (2005) – whose semantic head bears the predicationality feature, develops a similar relation with its direct or indirect objects as it has with the subject. Therefore, in Curteanu et al. (2007a), two different Argument - Predicate relations were defined, similar to the one between the subject and the verb, having as actors the verb and its direct and indirect object. The relations were called “redefined classical predications”, being instantiated by the predicational verb endowed with clitics as affixed inflexions (see figure 2.2). The clitics in Romanian are mandatory when their valence-\footnote{The extension of the Subject-Predicate relation to Argument-Predicate has been defined for Romanian, but the relations can be applied to other languages also. Even if the agreement between the predicate and its arguments is not lexicalized through clitics in the syntactic surface representation of the sentence, the semantic relation between the predicate and arguments exists nevertheless.}
commanded arguments are personalized or focused, disregarding the doubling (example (10)) or not (example (11)) of their corresponding semantic arguments - the direct or indirect object. In example (12), the sentence is grammatically incorrect\(^7\), since the clitic 1- is missing, while the direct object pe Mihai is lexicalized in the sentence. These observations indicate that the agreement, which was considered the specificity of the Subject - Predicate relation, exists also in the other Argument - Predicate relations, taking the form of clitic pronouns:

(10) \[\begin{align*}
&\text{Maria} &\text{a văzut pe Mihai.} \\
&\text{Maria} &\text{him-CL saw Mihai.} \\
&\text{“Maria saw Mihai.”} \\
\end{align*}\]

(11) \[\begin{align*}
&\text{Maria} &\text{a văzut.} \\
&\text{Maria} &\text{him-CL saw.} \\
&\text{“Maria saw him.”} \\
\end{align*}\]

(12) \[\begin{align*}
&*\text{Maria a văzut pe Mihai.} \\
&\text{Maria} &\text{a văzut.} \\
&\text{“Maria saw Mihai.”} \\
\end{align*}\]

Thus, the classical predication pair corresponds to the subject semantic role of “actor” or “actant”, while the added “redefined classical predications” may be associate, valence-driven, to the semantic roles of “patient” and “receiver”. All these are commanded by the presence (or absence) of the PREDF feature and the valence.

In Figure 2.2, \(\text{SUBJ}, \text{OBJD}, \text{OBJI}\) represents the syntactic categories of subject, direct and indirect objects, respectively. The \textit{obliqueness} is the order of the syntactic arguments in terms of their distance (and links) to the verb: obliqueness 0 means the argument is situated the closest to the verb with regard to their positions in the sentence, while obliqueness 3 means the argument is the furthest from the verb. For Romanian, a SVO language\(^8\), the subject is considered to be the least oblique element (the closest to the verb). The PREDFVerb

\(^7\)The * before the sentence indicates that it is a grammatically incorrect sentence

\(^8\) In linguistic typology, subject-verb-object (SVO) is a sentence structure where the subject comes first, the verb second, and the object third. Languages may be classified according to the dominant sequence of these elements. It is the second most common order found in the world, after SOV.
## 2. WHAT ARE SEMANTIC ROLES?

2. WHAT ARE SEMANTIC ROLES?

| (SUBJ\textsubscript{Obliqueness} \text{=} 0, PRED\textsubscript{Verb} | [\text{VG Tense Aspect}] |
| \text{Semantic Diathesis}(\text{SUBJ, OBJD, OBJI}) = (\theta(\text{SUBJ}), \theta(\text{OBJD}), \theta(\text{OBJI})) |
| \text{Agreement}(\text{SUBJ, Inflection VG}) |

| (OBJD\textsubscript{Obliqueness} \text{=} 1, PRED\textsubscript{Verb} | [\text{VG Tense Aspect}] |
| \text{Semantic Diathesis}(\text{SUBJ, OBJD, OBJI}) = (\theta(\text{SUBJ}), \theta(\text{OBJD}), \theta(\text{OBJI})) |
| \text{Agreement}(\text{OBJD, CliticOBJD VG}) |

| (OBJI\textsubscript{Obliqueness} \text{=} 2, PRED\textsubscript{Verb} | [\text{VG Tense Aspect}] |
| \text{Semantic Diathesis}(\text{SUBJ, OBJD, OBJI}) = (\theta(\text{SUBJ}), \theta(\text{OBJD}), \theta(\text{OBJI})) |
| \text{Agreement}(\text{OBJI, CliticOBJI VG}) |

**Figure 2.2: Redefined Predication** equally treating the Subject (Actant) and Objects (Arguments) relation to the predicational verb (Curteanu et al., 2007a)

indicates a verb which has the predicational feature. The Agreement function establishes anaphoric local bindings between the verbal group inflection and its subject on one hand, and between the syntactic OBJD and OBJI arguments and their pronominal clitics, affixed or not to the verbal group, on the other hand.

The Semantic\_Diathesis is not an elementary (atomic) feature, but a shallow function, taking as input the verbal group’s syntactic diathesis and the surface syntactic arguments order, and returning the verbal group’s semantic diathesis and the list of reordered syntactic arguments, having assigned the semantic roles of Actant, Patient or Addressee. Since the semantic roles are also called theta roles, figure 2.2 uses a theta function that maps the syntactic role to the semantic one: $\theta(\text{SUBJ}), \theta(\text{OBJD}), \theta(\text{OBJI})$. The semantic diathesis function first assigns the semantic diathesis to the verbal group, using the diathesis transformation method for the Romanian language presented in Curteanu and Trandabăț (2007), mapping the syntactic voices (active, passive or reflexive) into fine-grained semantic voices (active, passive, reflexive, reciprocal, impersonal, dynamic\(^9\)). In a second step, using the semantic diathesis, it reorders the syntactic arguments and assigns semantic labels. Thus, for example, if the verb is in an active semantic voice, the order of the syntactic argument is kept, and the SUBJ syntactic argu-

\(^9\)The list of fine-grained semantic diathesis was inspired by Irimia (1997).
ment receives the semantic Actant role, the OBJD receives the Patient role and the OBJI the Addressee role. If the verb is in a passive semantic voice (regardless of the syntactic voice), the SUBJ gets the Patient and the OBJI gets the Actant semantic role (in the passive diathesis, the semantic subject is represented syntactically as an object introduced with the preposition by). This solution forces the subject-actor and the subject-least_oblique_element (or grammatical subject) to take each one its own right place, in the right ordering.

2.1.5 Valence

The resources that contain descriptions of semantic frames and roles present the valence of predicational words in annotated contexts, in order to account for the link between the semantic roles and syntactic functions. The valence represents the combinatorial capacity of the verb, thus the property to open empty slots for the constituents of the sentence\textsuperscript{10}. Being closely related to predicationality, valence sets the number and the semantic and grammatical functionality of the constituents that the verb accepts (optional roles) or demands (mandatory roles). The most active and important valence demander is the verb, but predicational nouns or adjectives can also have valence if they contribute to form the semantic nucleus of the sentence:

\begin{enumerate}
  \item John [exposed]\textsubscript{PREDF} the situation.
  \item John’s [exposure]\textsubscript{PREDF} of the situation was very clear.
  \item The [exposed]\textsubscript{PREDF} situation was simple.
\end{enumerate}

In the communication process, predicational words act as nucleus, each having the possibility to ask different arguments in order to complete their sense. The degree of necessity of a specific constituent type for semantic plenitude varies, each verb allowing for a set of mandatory semantic roles, called arguments, and a set of optional, circumstantial semantic constituents, named adjuncts. The adjuncts are not specific to a verb, describing the context of the process rather than the process itself:

\textsuperscript{10}This property of predicational words to “demand” a specific argument structure is referred in the literature as either valence, valency or arity. The notion we will be use in this thesis is valence.
2. WHAT ARE SEMANTIC ROLES?

Verbs that only demand one, mandatory, argument:

\[[John]_{ARG1} \text{is}^{−PRED} \text{leaving}^{+PRED} \].

Verbs that demand two arguments:

\[[John]_{ARG1} \text{reads}^{+PRED} \text{a book}_{ARG2} \].

Verb with three arguments:

\[[John]_{ARG1} \text{gives}^{+PRED} \text{Mary}_{ARG2} \text{a book}_{ARG3} \].

Adjuncts examples for the sentences above are:

\[[John]_{ARG1} \text{is}^{−PRED} \text{leaving}^{+PRED} \text{today}_{ADJ1} \text{in a trip}_{ADJ2} \text{for Paris}_{ADJ3} \].

\[[John]_{ARG1} \text{reads}^{+PRED} \text{a book}_{ARG2} \text{in the library}_{ADJ1} \].

We can relate arguments to Tesnière’s actants and adjuncts to circumstantial phrases, as being idiosyncratic patterns of verbs encapsulated into a predicate-argument structure. Tesnière (1966) first made a distinction between the typical cases of a verb – actants – and its optional circumstantial modifiers, each verb attracting a definite number of actants, corresponding to its valence.

2.1.6 Case Grammar

Based on the linguistic knowledge representation introduced by Chomsky (1965), Ch. Fillmore introduces in 1968 the Case Grammar, an attempt to establish a semantic grammar in an era when most grammarians take syntax as the starting-point of any language theory. The starting point of his theory is dividing the language representation into two structures: Surface Structure (the syntactic knowledge) and Deep Structure (the semantic knowledge). The language process begins at the Deep Structure level with a non-verbal representation (an idea or a thought) and ends in the Surface Structure, as we express ourselves.

In case grammar, the semantic roles of the predicates were considered crucial to the characterization of verbs and clauses. Fillmore (1968) defines Case Frames as “characterizing a small abstract ‘scene’ or ‘situation’, so that to understand the semantic structure of the verb it was necessary to understand the properties of such schematized scenes”. According to Fillmore, the linguistic knowledge is organized around the predication, each predicational word (verb, noun or adjec-
2.1 Linguistic Background

tive) selecting a certain number of deep cases, some obligatory, others optional, which form its semantic frame. Thus, a case frame describes important aspects of semantic valence of verbs, adjectives and nouns.

The Case Notions (or Case Roles) are representations of the lexical arguments of a predicate at semantic level. Using a modified form of valence theory, Fillmore suggests that the verb establishes a set of cases in a sentence, which can be seen as slots that are waiting to be filled with syntactic structures. Thus, the case roles\textsuperscript{11} are defined as the role that the constituents of a sentence play in the interpretation of that sentence, these roles being defined in a fixed repertory. This inventory of Cases is a whole of universal concepts, possible innate, sufficient for the classification of the verbs of a language and reusable in all the languages. The list of Fillmore Cases first included six cases:

- The \textbf{Agentive} or the \textbf{Agent} – the Case of the entity (typically animated) perceived as the instigator of the action identified by the verb, the “doer of the action”:

  \[\text{Columbus}_{Agent} \text{ discovered America}.\]

- The \textbf{Instrumental} or the \textbf{Instrument} - the Case of the inanimate object or force implied in a causative way in the state or the action identified by the verb:

  \[\text{The window was broken [with a hammer]}_{Instrument}.\]

- \textbf{Dative} - the (animated) Case which is affected by the state or the action identified by the verb.

  \[\text{John gave the book [to Mary]}_{Dative}.\]

Although the name of this semantic case is similar to a syntactic case, a constituent in the semantic \textit{Dative} case does not have to be also in the syntactic \textit{Dative} case, as the example above may suggest. A counterexample is given in the sentence below, where \textit{Mary} has the \textit{semantic Dative} case, but the syntactic \textit{Accusative} case.

\textsuperscript{11}or deep cases, to be differentiated from surface, syntactic cases
2. WHAT ARE SEMANTIC ROLES?

John kissed [Mary].

This case name similarity led Fillmore to divide the Dative case into three other Cases in his later works: *Experiencer*, *Object* and *Goal*. These cases are explained later in this Section.

- **Factive** or **Resultative** – the Case of the entity or the object resulting from the action or the state identified by the verb, or identified as constituting the verbal sense. This role is later merged to the Object role.

  John damages [the table]_{Resultative}.

- **The Objective** or **Object** – the most semantically neutral Case, the Case of any named entity whose role in the action or the state identified by the verb is given by the semantic interpretation of the verb itself, the recipient of the verb action:

  [Our house]_{Object} burned down.

- **Locative** – the location or spatial orientation of the state or action identified by the verb. Fillmore distinguishes two types of Locative: the *Modal Locative*, as the Locative which does not belong to the verbal phrase, acting as an adjunct, and the *Propositional Locative*, the Locative which is a part of the verbal phrase, acting as an argument.

  John washes his car [in the garage]_{ModalLoc.} (ML)
  Mary keeps its jewels [in a little box]_{PropLoc.} (PL)

While the verb *wash* has the valence two (someone washes something), the verb *keep* with the sense it has in the example above, requires three arguments (someone keeps something in some safe place). Thus, the locative role is an optional adjunct in the first case, while an argument in the second case. However, the fine difference between the two locative cases led Fillmore to reconsider the locative case in his later work (see below).

After the preliminary list of cases, Fillmore noted that additional cases are needed, and refined his list of cases in later works with:

- **Experiencer** – the entity implied in a psychological event or a mental state.
[The sharks]$_{Experiencer}$ smelled blood.

Psychological predications are expressed on the surface structure of the language by the verbs of feeling (e.g. to hear), emotion (e.g. to like) and cognition (e.g. to believe).

**Comitative** – the Case, typically animated, used to express the accompaniment.

The children are [with Mary]$_{Comitative}$.

The **Location** is fine-grained into three Cases: the **Path**, the **Source** and the **Goal**. The Location case is viewed as a spatial localization, the Path represents a route, while the Source and Goal cases represent the starting, respectively the ending point of the action or state invoked by the verb.

Lucky raced [across the lawn]$_{Path}$ [to the edge of the forest]$_{Goal}$.

Max goes [from Paris]$_{Source}$ [to London]$_{Goal}$.

The **Temporal** case – defining the moment at which an event takes place.

They dined [at 5 p.m.]$_{Temporal}$.

Fillmore provided linguistic evidence for his deep cases. He argued for example that the subjects of sentences (a), (b) and (c) below have different thematic roles, namely Agent, Instrument and Object respectively, on the basis of facts in (d), assuming exactly one role per argument Noun Phrase. The unacceptable (e) represents a failed attempt to assign both Agent and Instrument roles to the subject, while in (f) the Instrument role is assigned to two different dependents.

a. [John]$_{Agent}$ broke [the window]$_{Object}$.

b. [A hammer]$_{Instrument}$ broke [the window]$_{Object}$.

c. [The window]$_{Object}$ broke.

d. [John]$_{Agent}$ broke [the window]$_{Object}$ [with a hammer]$_{Instrument}$.

e. *[John]$_{Agent}$ and a [hammer]$_{Instrument}$ broke [the window]$_{Object}$.

f. *[A hammer]$_{Instrument}$ broke [the window]$_{Object}$ [with a chisel]$_{Instr.}$.
2. WHAT ARE SEMANTIC ROLES?

2.1.7 Semantic Role Theories

Fillmore (1968) defined six case roles: Agent, Instrument, Dative, Factive, Object and Location. In Fillmore’s original theory, semantic roles were called deep cases (see Section 2.1.6). Central hypothesis in his theory is that there is a direct relation between deep cases and grammatical functions such as subject and object. Fillmore’s later work on lexical semantics led to the conviction that a small fixed set of deep case roles was not sufficient to characterize the complementation properties of lexical items. He added Experiencer, Comitative, Location, Path, Source, Goal and Temporal, and then other cases. This ultimately led to the theory of Frame Semantics, which later evolved into the FrameNet project (see Section 2.3.2).

Another influential theory on the subject of semantic roles was introduced by Ray Jackendoff in 1990, who started with the set of thematic roles that was originally introduced by Gruber in 1965: Theme, Source, Goal and Agent. Jackendoff proposed several modifications and refinements to this role inventory, based on a new formalism which he called conceptual semantics. According to Jackendoff, the meaning of a linguistic expression can be represented by a conceptual structure, composed of conceptual constituents. A conceptual constituent comprises one or more atomic semantic primitives. For every major constituent in the syntactic structure there has to be a corresponding constituent in the conceptual structure. The mapping between syntactic and conceptual structures is governed by correspondence rules. Jackendoff points out that, despite the fact that phonology (phonetic form) and meaning (logical form) have always been treated as if they are derived from syntax, the correct hierarchy were to put conceptual structure at one end, phonological structure at the other, with syntax in the middle.

Dowty (1991) looked at semantic roles from a different perspective. He stated that “there is in fact a notable absence in consensus about what thematic roles are”. Dowty rejected the idea of assuming individual thematic roles for each verb, rather than verb-independent roles. In his theory, Dowty focused on the problem of argument selection. Argument selection deals with the principles that determine which semantic arguments of a verb are expressed by which grammatical relation (e.g., subject, object). Dowty argued that thematic roles should not be
treated as discrete categories, but more like prototypical concepts. He defined two proto roles.

The idea behind his proto-role approach is that there are really only two thematic-role-like concepts involved in argument selection, and these are 'cluster concepts,' not discretely defined ones. He calls them the Agent Proto-Role and the Patient Proto-Role, and considers them the only thematic categories on which linking principles are stated. Dowty also establishes an inventory of properties for each of these two proto-roles:

Contributing properties for the Agent Proto-Role:

• volitional involvement in the event or state
• sentience (and/or perception)
• causing an event or change of state in another participant
• movement (relative to the position of another participant)
• exists independently of the event named by the verb

Contributing properties for the Patient Proto-Role:

• undergoes change of state
• incremental theme
• causally affected by another participant
• stationary relative to movement of another participant
• does not exist independently of the event, or not at all

No single property is essential for either role. Instead, Dowty gives the following procedure: “Argument selection principle: In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of Proto-Agent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of Proto-Patient entailments will be lexicalized as the direct object.” (Dowty, 1991)
2. WHAT ARE SEMANTIC ROLES?

There is no attempt to find some unifying semantics behind the lists of properties of the two proto-roles. *Proto-Agent* and *Proto-Patient* are ‘cluster concepts’ or ‘higher-order generalizations about meanings’ that need not even be considered as part of the competence grammar. Instead, Dowty suggests that the argument selection principle acts as a default in the acquisition of lexical items.

A number of linguists have borrowed aspects of the proto-role idea into their models of competence grammar. Davis’ (2001) theory of linking between word meaning and syntax uses proto-role properties in a multiple inheritance type hierarchy in the Head-Driven Phrase Structure Grammar framework (Pollard and Sag, 1994). Each proto-role property is encoded in the lexicon as a type within a rich hierarchy of types and sub-types. Davis assumes two ’macro-roles’ called *Actor* and *Undergoer*, similar to the terminology of Van Valin (Van Valin, 2005).

A theory that influenced modern lexical resources, like VerbNet and PropBank, was developed by Beth Levin. Levin (1993) argues that syntactic frames are a direct reflection of the underlying semantics. She defined verb classes based on the ability of a verb to occur or not in pairs of syntactic frames that are in some sense meaning preserving (diathesis alternations). VerbNet is a lexical resource in which the original set of Levin classes has been further subdivided into additional subclasses which are more syntactically and semantically coherent.

2.2 Description of Semantic Roles

Traditionally, semantic roles are part of the linking theory\textsuperscript{12}, a grammatical theory that describes the interaction between syntax and semantics. Semantic roles are used to assign meaning to syntactic constituents. The central question in the linking theory is how these roles can be inferred from syntax. What is relatively new though, is the automatic assignment of roles based on syntactic information. Today’s modern parsers are able to extract syntax from text accurately, which might have contributed to the renewed interest in semantic role labeling in recent

\textsuperscript{12}The relationship between such surface manifestations and semantic roles is the subject of linking theory (Levin and Rappaport Hovav, 2005) give a synthesis of work in this area). In general, linking theory argues that the syntactic realization of arguments of a predicate is predictable from semantics. Exactly how this relationship works is the subject of much debate.
2.2 Description of Semantic Roles

years. Historically, two types of semantic roles have been studied: Abstract roles such as Agent, Patient and Instrument, and roles more specific to a certain verb or class of verbs, like for instance Eater for the verb eat, or Seller for the verb sell.

2.2.1 Common List of Semantic Roles

In the simplest form of lexical semantic representation (Dowty, 1991) only two roles Proto-Agent and Proto-Patient are defined, but most theories define at least the six basic roles described by the Case Grammar (see Section 2.1.6). Here is a list of the major thematic relations usually considered:

**Agent:** deliberately performs the action

\[[Bill]_{Agent} \text{ ate his soup quietly.}\]

**Experiencer:** receives sensory or emotional input

\[\text{The smell of lilies filled [Jennifer’s]_{Experiencer} nostrils.}\]

**Theme:** undergoes the action but does not change its state. Sometimes used interchangeably with Patient

\[\text{I like [Kim]_{Theme}.}\]

**Patient:** undergoes the action and has its state changed

\[\text{The falling rocks crushed [the car]_{Patient}.}\]

**Instrument:** used to carry out the action

\[\text{Jamie cut the ribbon [with a pair of scissors]_{Instrument}.}\]

**Force or Natural Cause:** mindlessly performs the action

\[\text{[An avalanche]_{Force} destroyed the ancient temple.}\]

**Location:** where the action occurs

\[\text{Johnny and Linda played carelessly [in the park]_{Location}.}\]

**Direction or Goal:** where the action is directed towards

\[\text{The caravan continued on [toward the distant oasis]_{Direction}.}\]
2. WHAT ARE SEMANTIC ROLES?

**Recipient:** a special kind of goal associated with verbs expressing a change in ownership, possession.

I sent [John]$_{Recipient}$ the letter.

**Source:** where the action originated

The rocket was launched [from Central Command]$_{Source}$.

**Time:** the time at which the action occurs

The rocket was launched [yesterday]$_{Time}$.

**Beneficiary:** the entity for whose benefit the action occurs

I baked [Reggie]$_{Beneficiary}$ a cake.

**Manner:** the way in which an action is carried out

[With great urgency]$_{Manner}$, Agatha phoned 911.

**Purpose:** the reason for which an action is performed

Agatha phoned 911 right away [in order to get some help]$_{Purpose}$.

**Cause:** what caused the action to occur in the first place; not for what, rather because of what

[Since Clyde was hungry]$_{Cause}$, he ate the cake.

There are no clear boundaries between these relations. For example, in "the hammer broke the window", some linguists treat hammer as an *Agent*, others as *Instrument*, while some others treat it as a special role, different from these, verb specific.

2.2.2 Characteristics of Semantic Roles

Semantic roles are considered to have the following main characteristics:

1. There is a small fixed set of semantic roles.
2. Semantic roles are atomic (generally one role does not subsume another).
3. Every argument of every verb is assigned a semantic role or another.
2.2 Description of Semantic Roles

4. Each argument of a verb is assigned exactly one semantic role.

5. Semantic roles are uniquely assigned within a verb (e.g., only one argument can be dubbed agent).

6. Semantic roles are non-relational (e.g., the presence of a patient role in a verb does not imply the presence of an agent role as well).

Each of these characteristics of thematic roles presents certain problems. First of all: a small fixed set of thematic roles has never been agreed on and it seems unlikely that this will change. Proposals range from just a few to hundreds of them. As two extreme cases, consider Dowty’s approach (Dowty, 1991) with just two semantic roles from which all non-local values would derive, contrasted with the view in HPSG where each verb would assign its own peculiar semantic roles, different from the semantic roles of any other verb (Pollard and Sag, 1994).

A second problem that has been discussed in the literature extensively is the assumption that an argument is assigned exactly one role. Uniqueness does not seem to hold with animate subjects of verbs of motion in sentences such as the following:

(14) John ran into the house.

Gruber (1965) and Jackendoff (1990) claimed that John is both Agent, since it initiates and sustains the movement, and Theme, since it is the object that moves.

Distinctness (every argument of every verb is distinguished from the other arguments by the role it is assigned) is hard to establish in examples like the ones below:

(15) a. John played with Mary.
    b. John resembles his mother.
    c. A is similar to B.

In the examples above, both participants seem to be playing the same role.

Another problem is how and where to establish the boundary between role types. An example is in the case of direct objects that can be either Instruments (Instr.) or Comitatives (Com.).

(16) a. John burgled the house [with an accomplice]_{Instr./Com.}.
2. WHAT ARE SEMANTIC ROLES?

b. John won the appeal [with a highly-paid lawyer]_{Instr./Com.}.

2.2.3 Types of Semantic Roles

The Berkeley FrameNet project (Baker et al., 1998) is creating an on-line lexical resource for English, based on Frame Semantics and supported by corpus evidence. The key concept in the FrameNet method of annotation is a semantic frame, described as a representation of an object, event or situation. Each frame has its own set of roles. For example, the roles defined for the frame Research are field, question, researcher and topic. Frames are evoked by verbs that are semantically related to the frame. For example, the frame Research is evoked by investigate and research. Roles in FrameNet are called frame elements (FEs), the frame-evoking words are called lexical units (LUs).

Frame elements (or semantic roles) are classified in terms of how central they are to a particular frame (Ruppenhofer et al.). Three levels of semantic roles can be distinguished:

1. core elements
2. peripheral elements
3. extra-thematic elements

Core elements instantiate required roles and make the frame unique from other frames. Peripheral or Extra-thematic roles mark circumstantial notions such as Time, Place and Manner, i.e. they act as modifiers. They do not uniquely characterize a frame and are not mandatory. An example of an extra-thematic frame element is the role Reason in the sentence:

\[(17)\] [He]_{Wearer} would put on [a white overall]_{Clothing} [for the occasion]_{Reason}.

In the more syntactic-oriented approaches, such as Propbank, roles are typically divided into two categories: arguments, which capture a core relation, and adjuncts, which are less central. These are the terms that we will be using in this presentation. These distinctions carry over into Semantic Role Labeling systems, where we see that systems generally perform better on the more central arguments, and from the adjuncts, the location and time are the easiest to classify.
2.3 Semantic Role Resources

The intuition that semantic analysis can make a positive contribution to language-based applications has motivated the development of a number of lexical-semantic resources. Prominent among them are PropBank and FrameNet. The potential contribution of these resources is constrained by the information they contain and the level of effort involved in their development.

2.3.1 VerbNet

Previous research into the linking between semantic roles and syntactic realization is due mainly to the comprehensive study of Levin (1993). Levin argues that the syntactic frames are a direct reflection of the underlying semantics; the sets of syntactic frames associated with a particular Levin class reflect underlying semantic components that constrain allowable arguments. On this principle, Levin defines verb classes based on the ability of the verb to occur or not occur in pairs of syntactic frames that are in some sense meaning preserving (diathesis alternations). The classes also tend to share some semantic component.

VerbNet\footnote{VerbNet web address: http://verbs.colorado.edu/ kipper/verbnet.html} extends Levin’s classes by adding an abstract representation of the syntactic frames for each class with explicit correspondences between syntactic positions and the semantic roles they express, as in Agent REL Patient, or Patient REL into pieces for break.

\begin{align*}
(18) & \quad \text{a. John broke the window.} \\
        & \quad \text{b. The window broke into pieces.}
\end{align*}

The original Levin classes constitute the first few levels in the hierarchy, with each class subsequently refined to account for further semantic and syntactic differences within a class. The argument list consists of thematic labels from a set of 20 possible such labels (Agent, Patient, Theme, Experiencer, etc.). The syntactic frames represent a mapping of the list of thematic labels to deep-syntactic arguments. Additional semantic information for the verbs is expressed as a set (i.e., conjunction) of semantic predicates, such as motion, contact, transfer information, etc. Currently, all Levin verb classes have been assigned thematic labels.
2. WHAT ARE SEMANTIC ROLES?

and syntactic frames and over half the classes are completely described, including their semantic predicates. In many cases, the additional information that VerbNet provides for each class has caused it to subdivide, or use intersections of Levin’s original classes, adding an additional level to the hierarchy.

2.3.2 FrameNet

FrameNet (FN) is a lexicographic research project which produced a lexicon containing very detailed information about the syntax – semantics relations of the English predicational words (verbs, nouns and adjectives). This lexicon can be used both by human and computer users, becoming therefore an important lexical database for application in the natural language processing, lexical semantics, etc.

The Berkeley FrameNet\textsuperscript{14} is a lexical resource for the contemporary English, based on Frame Semantics, a linguistic theory that describes the conceptual structure of the linguistic meaning. The basic unit of analysis is the semantic frame, a “script-like structure of inferences, linked by linguistic convention to the meanings of the lexical units”, defined as a type of event or state in Fillmore (1985).

Each frame identifies a set of constituents, frame elements, which defines it and a set of lexical units which participate at its actualization. A lexical unit is a word or a sense of a polysemantic word for which combinatorial proprieties are defined. The description of a lexical unit in terms of Frame Semantics identifies the frames that forms a unique meaning and specifies the way that semantic roles are realized inside several structures dominated by the target word.

The key features of Berkeley FrameNet are automatically derived generalizations about frame structure and grammatical organization using corpus and the representation of the valences of the target words using frame semantics. The main corpus used by FrameNet is the British National Corpus\textsuperscript{15} (BNC), which contains more than 100.000.000 words from different sources (editorials, books, novels, advertising etc.) defining a great variety of the English written and spoken language.

The Berkley FN has recorded different type of information on separate anno-

\textsuperscript{14}FrameNet web address: http://framenet.icsi.berkeley.edu

\textsuperscript{15}BNC web address: http://info.ox.ac.uk/bnc
2.3 Semantic Role Resources

...tation layers: FE (Frame Element); GF (Grammatical Function); and PT (Phrase Type) (see Baker et al., 1998). The separation of layers makes it possible to represent many complex situations, such as when the constituent that realizes one frame element is contained within the constituent that realizes another, or when the semantic and syntactic constituency doesn’t match.

Since FN is primarily lexicographic, there was not a whole annotation of the BNC corpus; only a set of sentences which exemplify the range of combinatorial possibilities of a lexical unit, including all the types of syntactic constituents which can embody the frame elements were considered. The results are available in a set of XML documents, each containing the selected sentences annotated for the three levels.

2.3.3 PropBank

The Proposition Bank project\textsuperscript{16} takes a practical approach to semantic representation, adding a layer of predicate-argument information, or semantic role labels, to the syntactic structures of the Penn Treebank. The resulting resource covers every instance of every verb in the corpus and allows representative statistics to be calculated. The Proposition Bank (PropBank) provide a broad-coverage hand annotated corpus of such phenomena, enabling the development of better domain-independent language understanding systems, and the quantitative study of how and why these syntactic alternations take place. The PropBank focuses on the argument structure of verbs, and provides a complete corpus annotated with semantic roles, including roles traditionally viewed as arguments and as adjuncts.

Palmer et al. (2005) discuss the criteria used to define the sets of semantic roles to be annotated, and analyze the frequency of syntactic/semantic alternations\textsuperscript{17} in the corpus. The Proposition Bank objective is not a theoretical account of how and why syntactic alternation takes place, but rather to provide a useful level of representation and a corpus of annotated data to enable empirical study of these issues. While lexical resources such as Levin’s classes (Levin, 1993) and VerbNet provide information about alternation patterns and their semantics, the frequency

\textsuperscript{16}PropBank web address: http://www.cs.rochester.edu/ gildea/PropBank/Sort/

\textsuperscript{17}Syntactic/semantic alternation refer to the linking between semantic roles and syntactic realization, based on Levin’s (Levin, 1993) affirmation that the syntactic frames are a direct reflection of the underlying semantics.
2. WHAT ARE SEMANTIC ROLES?

of these alternations and their effect on language understanding systems has never
been carefully quantified. More recent work has attempted to group verbs into
classes based on alternations, usually taking Levin's classes as a gold standard
(e.g. Schulte im Walde and Brew, 2002). But without an annotated corpus of
semantic roles, this line of research has not been able to measure the frequency of
alternations directly, or, more generally, to ascertain how well the classes defined
by Levin correspond to real world data.

Because of the difficulty of defining a universal set of semantic or thematic
roles covering all types of predicates, PropBank defines semantic roles on a verb
by verb basis. The semantic arguments of an individual verb are numbered, from
0 to 4. For a particular verb, Arg0 is generally the argument exhibiting features
of an Agent, while Arg1 is a Patient or Theme. As examples of verb-specific
numbered roles, Palmer et al. (2005) give entries for the verbs accept and kick.
An important note is that the arguments are marked as numbered arguments
in the text, for generalization reasons, but their correspondence to verb-specific
roles is also provided (Arg0 is Acceptor for the verb accept and Kicker for the
verb kick).

Frameset accept.01 '‘take willingly’’

Arg0: Acceptor
Arg1: Thing accepted
Arg2: Accepted-from
Arg3: Attribute

Ex: [He]\Arg0 [would]\ArgM\MOD [n’t]\ArgM\NEG accept [anything of value]\Arg1
[from those he was writing about]\Arg2.

Frameset kick.01 ‘‘drive or impel with the foot’’

Arg0: Kicker
Arg1: Thing kicked
Arg2: Instrument (defaults to foot)

Ex1: [But]\ArgM\DIS [two big New York banks]\Arg0 seem [*trace*]\Arg0 to have
kicked [those chances]\Arg1 [away]\ArgM\DIR, [for the moment]\ArgM\TMP, [with the
embarrassing failure of Citicorp and Chase Manhattan Corp. to deliver $7.2
billion in bank financing for a leveraged buy-out of United Airlines parent UAL
2.3 Semantic Role Resources

Ex2: [John]$_{Arg0}$ tried [*trace*]$_{Arg0}$ to kick [the football]$_{Arg1}$, but Mary pulled it away at the last moment.

In addition to verb-specific numbered roles, PropBank defines several more general roles that can apply to any verb. These adjuncts (circumstantial objects) are marked as ARG-Ms (modifiers). They can appear in any verb’s frame, valence-independently. Figure 2.3 gives the list of PropBank annotated adjuncts.

<table>
<thead>
<tr>
<th>ArgM type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIR</td>
<td>direction</td>
</tr>
<tr>
<td>LOC</td>
<td>location</td>
</tr>
<tr>
<td>MNR</td>
<td>manner</td>
</tr>
<tr>
<td>TMP</td>
<td>time</td>
</tr>
<tr>
<td>EXT</td>
<td>extent</td>
</tr>
<tr>
<td>REC</td>
<td>reciprocal</td>
</tr>
<tr>
<td>PRD</td>
<td>secondary predication</td>
</tr>
<tr>
<td>PNC</td>
<td>purpose</td>
</tr>
<tr>
<td>CAU</td>
<td>cause</td>
</tr>
<tr>
<td>DIS</td>
<td>discourse connectives</td>
</tr>
<tr>
<td>ADV</td>
<td>general-purpose adverbs</td>
</tr>
<tr>
<td>MOD</td>
<td>modal verb</td>
</tr>
<tr>
<td>NEG</td>
<td>negation marker</td>
</tr>
<tr>
<td>DSP</td>
<td>direct speech</td>
</tr>
<tr>
<td>RLC</td>
<td>relative clauses</td>
</tr>
</tbody>
</table>

Figure 2.3: PropBank list of annotated adjuncts with their explanations

In some cases, an argument may span over different parts of a sentence, the label C-Arg is used to specify the continuity of the arguments, as shown in the example below:

(19) [The pearls]$_{ARG1}$, [I]$_{ARG0}$ [said]$_{TARGET}$, [were left to my daughter-in-law]$_{C-ARG1}$.

Moreover in some cases, an argument might be a relative pronoun that in fact refers to the actual agent outside the clause. In this case, the actual agent is labeled as the appropriate argument type, as the pearls is in the next example, while the relative pronoun is instead labeled as R-Arg. For example:
2. WHAT ARE SEMANTIC ROLES?

(20) [The pearls]$_{ARG1}$, [which]$_{R-ARG1}$ [I]$_{ARG0}$ [left]$_{TARGET}$ [to my daughter-in-law]$_{ARG2}$, are fake.

NomBank (Meyers et al., 2004a,b) is an annotation project at New York University that is related to the PropBank project at the University of Colorado. The goal of the NomBank project is to mark the sets of arguments that co-occur with nouns in the PropBank Corpus (the Wall Street Journal Corpus of the Penn Treebank), just as PropBank records such information for verbs. These resources help the user to label the various arguments and adjuncts of the head nouns with roles (sets of argument labels for each sense of each noun). The two project teams are making a coordinated effort to insure that, when possible, role definitions are consistent across parts of speech. For example, PropBank’s frame file for the verb decide is used in NomBank annotation of the noun decision (Gerber et al., 2009).

2.3.4 FrameNet vs. PropBank

The PropBank project and the FrameNet project share the goal of documenting the syntactic realization of arguments of the predicates of the general English lexicon by annotating a corpus with semantic roles. Despite the two projects’ similarities, their methodologies are quite different. FrameNet is focused on semantic frames, which are defined as a schematic representation of situations involving various participants, props, and other conceptual roles. The project methodology has proceeded on a frame-by-frame basis, that is by first choosing a semantic frame (e.g., Commerce), defining the frame and its participants or frame elements (Buyer, Goods, seller, Money), listing the various lexical predicates which invoke the frame: buy, sell, etc., and then finding example sentences of each predicate in a corpus (the British National Corpus was used) and annotating each frame element in each sentence. The example sentences were chosen primarily to ensure coverage of all the syntactic realizations of the frame elements, and simple examples of these realizations were preferred over those involving complex syntactic structure not immediate relevant to the lexical predicate itself. Only sentences where the lexical predicate was used “in frame” were annotated. A word with multiple distinct senses would generally be analyzed as belonging to different frames in each sense, but may only be found in the FrameNet corpus in
the sense for which a frame has been defined. The semantic frames are a helpful way of generalizing between predicates; words in the same frame have been found frequently to share the same syntactic argument structure (Gildea and Jurafsky, 2002b).

In contrast with FrameNet, PropBank is aimed at providing data for training statistical systems and has to provide an annotation for every clause in the Penn Treebank, no matter how complex or unexpected. Similarly to FrameNet, PropBank also attempts to label semantically related verbs consistently, relying primarily on VerbNet classes for determining semantic relatedness. However, there is much less emphasis on the definition of the semantics of the class that the verbs are associated with, although for the relevant verbs additional semantic information is provided through the mapping to VerbNet. The PropBank semantic roles for a given VerbNet class may not correspond to the semantic elements highlighted by a particular FrameNet frame.

PropBank has recently started to address also nouns, beside verbs, whereas FrameNet included from the start also nouns and adjectives. PropBank annotation also differs in that it takes place with reference to the Penn Treebank trees – not only are annotators shown the trees when analyzing a sentence, they are constrained to assign the semantic labels to portions of the sentence corresponding to nodes in the tree. Parse trees are not used in FrameNet; annotators mark the beginning and end points of frame elements in the text, and add a grammatical function tag expressing the frame element’s syntactic relation to the predicate.

### 2.3.5 Combining Semantic Frames with other Lexical Resources

Recently, there has been a growing interest in more in-depth semantic analysis for practical NLP tasks, in particular as a basis for open-domain information access. Large-scale lexical semantic resources, such as WordNet, have been developed and put to use for approximate semantic modeling in many applications. The FrameNet and PropBank projects are developing lexical semantic resources that focus on the modeling of predicate-argument structure.

Authors have begun wondering about the combinatorial possibilities of different resources. Terenzi and Eugenio (2003) report in their work presenting the
2. WHAT ARE SEMANTIC ROLES?

attempt to automatically build a corpus of instructional text annotated with lexical semantics information, the integration of two lexical resources, Verb-Net for verbs and CoreLex for nouns. They show that a parser which uses such a lexicon and ontology performs extremely well on instructional text.

Burchardt et al. (2005) present a rule-based system for the assignment of FrameNet frames by way of a “detour via WordNet”. The system can be used to overcome sparse-data problems of statistical systems trained on current FrameNet data. The Berkeley FrameNet database groups words and expressions (lexical units, LUs for short) into semantic classes (frames) and lists semantic roles for each frame. This type of lexical semantic information is particularly useful for information access tasks, like Information Extraction (IE), or Question Answering (QA). Burchardt, Erk and Frank (Burchardt et al., 2005) have investigated the use of FrameNet frames for building partial text meaning representations, in order to use them in applications like IE and QA as addressed e.g. by the Recognizing Textual Entailment (RTE) Challenges. Semantic representations building on frames provide normalizations over surface realizations (e.g. active/passive, verb/nominalization) and thus a sensible granularity for these applications.

Two major tasks in the (automatic) annotation of texts with frames are the frame assignment problem, i.e. the identification of the proper frame for a given lexical unit, and the semantic role assignment problem, i.e., the assignment of the frame’s semantic roles to major sentence constituents. As a base system for frame assignment, Burchardt, Erk and Frank are using a system that treats this task as (supervised) word sense disambiguation, using the FrameNet collection of annotated sentences as training data. But as FrameNet is still a growing resource, the annotation of contiguous text is confronted with two problems. The first is a problem of coverage: For example, for the 574 sentences of the RTE development corpus, with an average of 16.24 words/sentence, the WSD system yields an assignment of 2.7 frames and 3.6 frame elements per sentence.

For example, the verb **skim** is listed as a lexical unit for four frames: (All examples are from the FrameNet corpus, some are abbreviated.)

**READING:** Skimming a chapter for its main idea may be done over coffee.

**REMOVING:** Remove the vanilla pod, skim the jam, and let it cool.

**SCRUTINY:** She skimmed through the newspaper clippings.
2.3 Semantic Role Resources

**SELF\_MOTION**: We skimmed across the surface of that sodding lake whilst all around us gathered the dark hosts of hell.

The frames Reading, Removing and Self\_Motion constitute clearly distinguished senses of skim. Reading and Scrutiny are hard to distinguish as far skim is concerned, even though in general they describe different situations, with different semantic roles and different LUs: Reading includes devour, peruse while Scrutiny has LUs like analyze, search, survey. The same holds for frame elements (semantic roles).

The second problem concerns lacking senses in the current FrameNet resource and is caused by the fact that FrameNet is being constructed one frame at a time, rather than one lemma at a time. While a lack of coverage leads to missing frame assignments, lacking senses result in wrong assignments. Therefore, they propose the use of WordNet as a “detour to FrameNet”. They use WordNet synsets as an interface layer to propose LU-frame pairs that are missing in the FrameNet database. They employ a WordNet-based WSD system to annotate lexical units in unseen texts with their contextually determined WordNet synset. Frame assignment can then proceed by using not just a single word, but also its synonyms and hypernyms.

This method exploits the fact that sense discrimination in WordNet is in general more fine-grained than FrameNet senses. However, in certain cases, close Wordnet relatives map to distinct FrameNet senses. Moreover, the method is dependent on the assignment of the correct synset by an independent WordNet-based WSD system. Thus, errors in the initial assignment of synsets immediately affect the quality of frame assignment.

**Shi and Mihalcea (2005)** try to create a system that integrates three different lexical resources: FrameNet, VerbNet, and WordNet, into a unified, richer knowledge-base, to the end of enabling more robust semantic parsing. The construction of each of these lexical resources has required many years of laborious human effort, and they all have their strengths and shortcomings. By linking them together, they build an improved resource in which (1) the coverage of FrameNet is extended, (2) the VerbNet lexicon is augmented with frame semantics, and (3) selectional restrictions are implemented using WordNet semantic classes.

The goal of a semantic parser is to identify semantic relations between words
2. WHAT ARE SEMANTIC ROLES?

in a text, resulting in structures that reflect various levels of semantic interpretation. Such structures can be used to improve the quality of natural language processing applications by taking into account the meaning of the text. Automatic techniques for semantic parsing have been successfully used in Information Extraction and Question Answering, and are currently evaluated in other applications such as Machine Translation and Text Summarization. The process of semantic parsing typically implies a learning stage, where the semantic structures to be identified are acquired from an existing lexical resource, which explicitly identifies the range of possible semantic relations between words in a text. While there are several lexical resources suitable for semantic parsing, built with extensive human effort over years of work - including FrameNet, VerbNet, WordNet, or PropBank - all previous approaches to semantic parsing have relied exclusively on only one of them, as there are no connections between these resources that would enable their exploitation in an unified way. However, each resource encodes a different kind of knowledge and has its own advantages.

Green et al. (2004) have created SemFrame, a system that induces frame semantic verb classes from WordNet and LDOCE. Semantic frame types of an intermediate granularity have the potential to fulfill an Interlingua role within a solution to the paraphrase problem\(^\text{18}\). Until now, semantic frames have been generated by hand, based on native speaker intuition; the FrameNet project now couples this generation with empirical validation. Only recently has this project begun to achieve relative breadth in its inventory of semantic frames. To have a comprehensive inventory of semantic frames, however, generating semantic frames semi-automatically is a necessity.

\(^{18}\)An important issue for computational linguists and lexicographers is the question of meaning-equivalent paraphrases, including lexical synonymy, conversives (buy/sell), idioms (kick the bucket/die), and more extended paraphrases, such as Its network of eighteen organizations has lent a billion dollars to micro-enterprises and The network comprises eighteen organizations which have disbursed a billion dollars to micro-enterprises.
2.3.6 Creating Semantic Role Resources for Languages other than English

In the most general terms, multilingual data refers to the existence of similar resources for more than one language. Practically, multilingual data can only be compared if it represents the same kind of linguistic information: e.g., the lexicon, orthography, phonology, syntax or semantics. At the particular level(s) of information each of them encode, a set of language resources can exhibit different degrees of relatedness. Two ways of relatedness are distinguished (Lönneker-Rodman, 2007): organizational similarity and interrelatedness. Organizational similarity is exhibited by multilingual language resources that use the same organizational principle. Resources covering the same kind of linguistic information and exhibiting the same organizational principle can further be interrelated. The interrelation, also called linking, mapping, or alignment, can be subdivided with respect to the way in which links are established. In the course of the EuroWordNet project (Vossen, 1999), which aimed to build an interrelated multilingual WordNet for eight European languages, the project team distinguished two methods, which they called the Merge approach and the Expand approach.

When the Merge approach is adopted, independent resources for different languages are first built from scratch. Later, links that relate selected types of components cross-linguistically are added. An important resource created using this approach is the Romanian Wordnet (Tufiş and Cristea, 2002), as part of the larger multilingual Balkanet project. An analysis of the trans-lingual preservation of the synsets between English and Romanian is given in Cristea et al. (2004).

In the opinion of Pianta et al. (2002), this approach potentially involves difficulties during the alignment phase. They argue that independently built resources might be more divergent than linguistically necessary because resource designers are free to make different decisions at a more abstract (world-knowledge related) level.

With the Expand approach, a resource for one language, which is regarded as stable at that time, is transferred to another language. This implies that initially, the overall structure of the resource is kept unchanged and only obviously language specific information is replaced. A problem for this method is the well-known fact that any given two languages might lexicalize different concepts, and
2. WHAT ARE SEMANTIC ROLES?

therefore also the relations between valid building blocks vary from language to
language. For this reason, the Expand approach requires a subsequent “cleaning”
of the resource resulting from the initial transfer.

The Expand approach tends to produce structurally highly similar resources. A problem is that this can entail the risk of neglecting language-specific differences in lexicalization and therefore in the structure of the lexicon. In particular, as a result of the translation of lexicalizations proper to Language A, “artificial” lexical labels (such as infrequent multi-word units) might appear in Language B. Also, unless the adaptation step is supported by further monolingual evidence from Language B, concepts lexicalized in this language alone (but not in A) are likely to be missed.

In summary, the Expand approach tends to produce structurally highly similar resources, at the risk of neglecting language-specific differences in lexicalization and, therefore, the structure of the lexicon.

Multilingual databases built along the FrameNet model are organizationally similar. In principle, FrameNets allow for cross-lingual relations at many levels, including those of frames, lexical units, lemmas, or even word forms and annotated sentences. The language specific aspects of definitions include example sentences and generalizations of syntactic realization patterns. The English examples are replaced in FrameNets for other languages by original examples from those languages that fulfill the same function; in other words, example sections are organizationally similar but not necessarily semantically equivalent.

2.3.6.1 German

The SALSA project (Erk et al., 2003), consists in tagging a German corpus manually and semi-automatically with semantic roles in order to derive a large domain independent lexical semantic resource. The corpus used is TIGER (Brants et al., 2002), a 1.5 Million word corpus of newspaper text with manually annotated syntactic structure, and the semantic annotation is performed using FrameNet frame semantic roles.

Each frame is annotated in the form of a flat frame of depth one. Its root is labeled with the frame, and its edges are anchored to nodes of the syntactic structure. The edge(s) leading to the word(s) that evoke the frame (the FEE or frame-evoking element) are unmarked. The edge(s) leading to frame semantic
role bearers (or frame elements) are marked with the semantic role. Frames are independent of each other.

All verbs and the deverbal nouns are tagged as frame-evoking, plus multi-word expressions. If a frame elements or frame evoking element consists of more than one node of the syntactic structure, the frame tree will have either one edge that is split, or two or more edges with the same label.

The developing of large role-annotated corpora raises another problem. To represent these corpora, a standardized multi-level annotation format, which integrates semantic role annotation with other linguistic annotation levels, is necessary. Erk and Pado present in 2004 two XML formats for the description and encoding of semantic role information in corpora: the TIGER/SALSA XML format, which provides a modular representation for semantic roles and syntactic structure, and the Text-SALSA XML format, a lightweight version of TIGER/SALSA XML, designed for manual annotation with an XML editor rather than a special tool. Due to its very general underlying model, TIGER/SALSA XML is limited neither to the TIGER corpus nor the task of annotating semantic roles. TIGER XML describes trees with arbitrary node and edge labels and crossing edges. It has been designed with the express purpose of being able to encode different linguistic frameworks. TIGER/SALSA XML can encode semantic roles that are verb-specific (as in PropBank), verb-group-specific (as in FrameNet), or general (as in the Prague Treebank).

2.3.6.2 Japanese

The Japanese FrameNet (JFN) research project, which started in July 2002, is headquartered on Hiyoshi Campus of Keio University and includes researchers from Keio University and University of Tokyo. The JFN corpus contains currently approximately 1 million sentences, taken from Kyoto University Annotated Text Corpus. Kyoto University Corpus contains morphologically and syntactically annotated data for 40,000 sentences (about 1.6 million words). A search tool has been developed in JFN, allowing the searches for both the root form and conjugated forms of a keyword at the same time, using the morphological annotations.
2. WHAT ARE SEMANTIC ROLES?

2.3.6.3 Spanish

Spanish FrameNet (SFN) is a research project sponsored by the Department of Science and Technology of Spain, and developed at the Autonomous University of Barcelona by a team of researchers led by Prof. Carlos Subirats. The goal of Spanish FrameNet (Subirats-Ruggeberg and Petruck, 2003) is to annotate corpus citations and to discover the valence patterns for a large number of predicational words, showing how those valence patterns are instantiated in actual sentences.

Each Spanish FrameNet entry will provide links to other lexical resources, including Spanish EuroWordNet synsets and syntactic subcategorization frames. The project’s deliverables will consist of the SFN database itself: lexical entries for individual word senses, frame descriptions, and annotated subcorpora. SFN uses a 300 million-word corpus, the same annotation software and database structure as that of the Berkeley project. The SFN Corpus includes both New World and European Spanish. It is composed of texts of different genres, primarily newspapers, newswire texts, book reviews, and humanities essays. These texts of various origins and genres make a grand total of 350 million words.

2.3.6.4 Adding Romanian to the Semantic Role Resources Map

Interests to realize semantic frames databases as a stable starting point in developing semantic knowledge based systems exists in countries such as Germany (the Salsa project), England (the PropBank project), United States (the FrameNet project), Spain, Japan, etc. One of the aims of this work is to create a semantic frame database for Romanian, similar to the ones existing for English. Since creating language resources demands many temporal, financial and human resources, the solution adopted by us regards the import of standardized annotation of a resource developed for a specific language to other languages. This approach will be discussed in Chapter 5.

2.4 Conclusions

The natural language processing community has recently experienced a growth of interest in semantic roles, since they describe WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW etc. for a given situation, since they contribute
to the construction of meaning. The semantic roles played by the participants in the denoted event or state, such as *Baker* and *Bakee* in the baking event, are also called thematic roles or theta-roles. This chapter has presented the linguistics underlying the semantic role theories, starting with the notions of perspective, predicationality and valence. After the first set of semantic roles introduced by Fillmore in 1965, reduced to their mere essence by Dowty’s proto-roles, an inventory of the most usual considered semantic roles was presented, with the main characteristics of thematic roles.

In the last decades, hand-tagged corpora that encode such information for the English language were developed. Three such corpora (VerbNet, FrameNet and PropBank) are discussed, with their resemblance and individualities. The chapter ends by presenting the attempts at creating semantic roles resources for languages other than English: German, Spanish, and Japanese. This thesis will present the first attempt to create a Romanian semantic role resource.
2. WHAT ARE SEMANTIC ROLES?
3

Approaches to Semantic Role Labeling

Researchers have concentrated their effort during many years to construct large manually annotated corpora. FrameNet for example, a subset of the British National Corpus, currently contains about 150,000 annotated sentences in about 960 frames. However, for role semantics to become relevant for language technology, robust and accurate methods for automatic semantic role assignment are needed. In recent years, a number of studies, such as (Chen and Rambow, 2003; Gildea and Jurafsky, 2002a), have investigated this task using as training datasets the FrameNet or PropBank corpora. Role assignment has generally been modeled as a classification task: a statistical model is created using manually annotated data and later used to assign a role label out of a fixed set to every semantic role of the predicates in a new, unlabeled sentence. The work on Semantic Role Labeling (SRL) has included a broad spectrum of probabilistic and machine-learning approaches, mostly supervised systems, using the large role-annotated resources.

The existing studies have used different statistical frameworks, but have largely converged on a common set of features to base their decisions on, namely syntactic information (path from predicate to constituent, phrasal type of constituent, etc.) and lexical information (head word of the constituent, predicate form or lemma, etc.).

With the SensEval-3 competition\(^1\) and the CONLL 2008 and CONLL 2009

\(^{1}\)SemEval web address: http://www.senseval.org/
3. APPROACHES TO SEMANTIC ROLE LABELING

Shared Tasks, Automatic Labeling of Semantic Roles (identifying frame elements within a sentence and tag them with appropriate semantic roles given a sentence, a target word and its frame) has become increasingly present among researchers worldwide.

Most general formulation of the SRL problem supposed determining the semantic label of (usually, but not always, contiguous) substrings (syntactic phrases) of the sentence $s$, given a target predicate $p$.

(1) $[\text{Bill}]_{\text{Agent}} [\text{ate}]_{\text{TARGET}} [\text{his soup}]_{\text{Theme}} [\text{quietly}]_{\text{Manner}}$.

(2) $[\text{By working hard}]_{\text{Theme}}, [\text{he}]_{\text{Agent}} [\text{said}]_{\text{TARGET}}, [\text{you can achieve a lot}]_{\text{Continued Theme}}$.

The most common errors of SRL systems are:

- overlapping argument strings - the system identifies two embedded role:
  $[\text{Bill}]_{\text{Agent}} [\text{ate}]_{\text{TARGET}} [[\text{his soup}]_{\text{Theme}} \text{ quietly}]_{\text{Manner}}$.

- repeated arguments - instead of one single role, the system annotates two roles with the same label:
  $[\text{Bill}]_{\text{Agent}} [\text{ate}]_{\text{TARGET}} [\text{his soup}]_{\text{Agent}} [\text{quietly}]_{\text{Manner}}$.

- missing arguments - the system cannot find a semantic role, and, since all constituents are supposed to have a role, one of the correctly identified roles (the Theme in this case) gets wrongly extended to cover the constituent that had no assigned role:
  $[\text{Bill}]_{\text{Agent}} [\text{ate}]_{\text{TARGET}} [\text{his soup quietly}]_{\text{Theme}}$.

We consider that, by introducing a set of restrictions and relaxing other constraints, these errors could be avoided. For the first error type, a constraint of no imbrications between roles can be envisaged. For the second error type, a

\footnote{ConLL web address: http://ifarm.nl/signll/conll/}
\footnote{The annotation of this example is simplified for exemplification of the problems that SRL systems have, hence the arguments for the verb achieve are not marked: $[\text{By working hard}]_{\text{Manner}}, \text{he said, [you}]_{\text{Agent}} \text{ can [achieve]}_{\text{TARGET}} [\text{a lot}]_{\text{Theme}}$.}
restriction of only one role type per predicate must be set (in the presented sen-
tence, only one Agent role can be assigned to a semantic role of the predicate\textsuperscript{4}. As for the third error type, we believe that the assumption that all constituents are supposed to have a role is too strong. This constraint holds for the SRL systems that only consider one predicational verb per sentence (as most of the SRL systems currently developed do), but when more predicational target words are considered, a constituent may be a semantic role for one predicate, while not being a semantic role for another constituent. If considering the example in sentence (2) above, we observe that for the target word said, the constituent he is an Agent, but for the target word achieve, he is not within the scope of the predication, thus cannot have a semantic role. These restrictions were considered in a post-processing step of our SRL system presented in Chapter 4.

3.1 Characteristics of Semantic Role Labelers

3.1.1 Types of Labeled Categories

The different approaches to semantic role labelers found in the literature can be divided into four categories, with respect to the type of tokens they classify:

- constituent-by-constituent (C-by-C);
- phrase-by-phrase (P-by-P);
- word-by-word (W-by-W);
- relation-by-relation (R-by-R).

In the C-by-C semantic role labeling, the syntactic tree representation of a sentence is linearized into a sequence of its syntactic constituents (non-terminals). Then each constituent is classified into one of several semantic roles using a number of features derived from the sentence structure or a linguistic context defined for the constituent token.

In the P-by-P and W-by-W methods, described in (Hacioglu and Ward, 2003), the problem is formulated as a chunking task and the features are derived for

\textsuperscript{4} with the exception of interrupted arguments, as Continued Theme - see Section 2.3.3.
each base phrase and word, respectively. The tokens are classified into one of the semantic labels using an IOB (inside-outside-begin) representation and a set of classifiers, one for each class.

The R-by-R method is based on dependency trees generated from constituency trees. Although these systems do not use more information than C-by-C systems, the information is structured in a different manner and, consequently, the nature of some linguistic features is quite different. Hacioglu (2004) point out that this restructured information is very useful in localizing the semantic roles associated with the selected predicate, since the dependency trees directly encode the argument structure of lexical units populated at their nodes through dependency relations.

### 3.1.2 Types of Vector Spaces

The basic hypothesis in applying machine learning techniques to semantic roles is that a frame semantic resource can be (at least partially) modeled and represented by a suitable semantic vector space model (VSM). The intuition is that a semantic space is a model able to capture the notion of frame (i.e. the property of “being characteristic of a frame”) for both lexical elements (lexical units, frame elements, etc.) and full sentences. This can be achieved by representing these elements as distributional vectors in the space - i.e. by using co-occurrence vectors.

Vector space models (VSM) are widely used in NLP for representing the meaning of words or other lexical entities (Pennacchiotti et al., 2008). The basic intuition is that the meaning of a target word is somehow defined by the context in which it appears (the Distributional Hypothesis introduced by Harris in 1964). The context can be defined in different ways: as the set of words surrounding the target word, as the paragraph in which it appears, the document, and so on. Vector spaces are used to model this intuition, by collecting statistics about the contexts of a target word within a large corpus.

Following this idea, a target word \( tw \) is represented by a vector, whose dimensions are the contexts in which it appears. When contexts are words \( w \) in an \( n \)-window of the target - i.e. \( tw \) co-occurs with the context word if \( w \) is within \( n \)-tokens on the left or on the right - we have a **word-based VSM**. When contexts are documents or sentences in which the target appears, we talk
about document-based VSM. The value of each dimension is given by the co-occurrence value of the target word with the given context. In the simplest case, these values are counts, i.e. the number of times that the target and the context co-occur in the corpus. More sophisticated association measures can be used to compute co-occurrence values. The most widely used are conditional probability ($p(w|tw)$) of the word given the target or point-wise mutual information ($pmi(tw,w)$) between the target and the word.

Computationally, a VSM is represented by a matrix, whose each row describes a target word with columns describing contexts. This matrix is used to calculate the distributional similarity between two targets, by computing the distance of their vectors. Different distances are here used expressing similarity measures, such as the cosine between vectors or their Euclidean Distance.

Using VSM for modeling a FrameNet-like resource poses a fundamental question: What is the relation between the geometry of a vector space model and the linguistic notion of frame? How is the distance in the space correlated with the notion of similarity between frame predicates? Pennacchiotti et al. (2008) proposes an effective space model for the lexical units that represent a frame. The goal is then to have similar vectors for LUs belonging to the same frame (or to related frames). For example, the lexical units killer and suicide are closer in the space, as they both evoke the Killing frame, with respect to killer and eat, as these latter express unrelated frames (Killing vs. Ingestion).

Three different types of spaces are most promising to model LUs: word-based, document-based and syntax-based. The latter type is similar to word-based VSM, differing on the fact that contexts are not co-occurring words but co-occurring syntactic relations. While all these spaces express distributional similarity, from a semantic perspective, they tend to model different types of relations.

Syntax-based spaces Syntax-based spaces are good at modeling semantic similarity. Two target words close in the space are likely to be close also in a is-a hierarchy (Budanitsky and Hirst, 2006), i.e. they are synonyms, antonyms, hyperonyms, etc. (e.g. human/man, dog/animal, good/bad). This is explained by the fact that contexts are syntactic relations, and then targets with the same Part of Speech are much closer than targets of different types. Experiments reported in (Pado and Lapata, 2007) support this claim. In other terms, syntax-based spaces tend to capture paradigmatic relations and
3. APPROACHES TO SEMANTIC ROLE LABELING

to disregard syntagmatic relations. According to Saussure (1916), paradigmatic relations relate two words likely to appear in the same context but not at the same time (*in absentia*). Syntagmatic relations stand between two words when they are likely to be combined together in the same texts (*in presentia*).

**Word-based spaces** These spaces model a more generic notion of semantic relatedness. Two targets close in the space are likely to be related by some type of generic semantic relation. The nature of relations captured by these spaces is matter of debate. In practice, word-based spaces have shown effective both to capture syntagmatic and paradigmatic relations. This is supported by the fact that words with similar co-occurrences can be both words occurring together in a text (e.g. doctor and patient in “the doctor operated the patient in the hospital”, sharing the same contexts *operated*) and substitution words (e.g. doctor and surgeon in “the (doctor/surgeon) operated the patient in the hospital”). Experiments in (Pado and Lapata, 2007) support this idea, showing that word-based spaces capture syntagmatic relations such as meronymy (door/house), conceptual association (doctor/hospital) and phrasal association (private/property), better than syntax-based spaces, while still capturing paradigmatic relations.

**Document-based spaces** Document spaces are historically used to model topic similarity, in the sense that words similar in the space tend to refer to the same topics. This definition has a strong Information Retrieval flavor: words close in the space are those occurring in the same documents, i.e. those focusing on a specific topic. This clearly depends on the nature of the corpus at hand, and on the choice of the context (document, paragraph or sentence). Topic similarity mainly involves co-occurring words (e.g. doctor/hospital for the medical topic). Document-based spaces should then better capture syntagmatic relations.

From our perspective, the notion of frame has both a syntagmatic and paradigmatic flavor. Prototypical situations (i.e. frames) involve different types of participants and facts that in the real world can either stage together or be one the substitute of the another (the **victim** and the **killer** in the first case, the **suicide** and the **killer** in the second case). In the same way, lexicalizations of
3.1 Characteristics of Semantic Role Labelers

Participants and facts can occur together in the text describing the situation or substitute one another. Then, in our particular setting, we focus on word-based and syntax-based spaces, as they seem to better capture this ambivalent notion.

3.1.3 Semantic Role Labelers Architecture

Given a sentence and a designated verb, the SRL task consists of identifying the boundaries of the arguments of the verb predicate (argument identification) and labeling them with semantic roles (argument classification). The most common architecture for automatic SRL, presented in (Marquez et al., 2008), consists of the following steps to achieve these subtasks:

**The first step** in SRL typically consists of filtering (or pruning) the set of argument candidates for a given predicate. Because arguments may be a continuous or discontinuous sequence of words, any subsequence of words in the sentence is an argument candidate. Exhaustive exploration of this space of candidates is not feasible, because it is both very large and imbalanced (i.e., the vast majority of candidates are not actual arguments of the verb). The simple heuristic rules, presented in (Xue and Palmer, 2004), are commonly used to perform filtering because they greatly reduce the set of candidate arguments, while maintaining a very high recall.

**The second step** consists of a local scoring of argument candidates by means of a function that outputs probabilities (or confidence scores) for each of the possible role labels, plus an extra “no-argument” label indicating that the candidate should not be considered an argument in the solution. In this step, candidates are usually treated independently of each other. A crucial aspect in local scoring is the representation of candidates with features, rather than the particular choice of classification algorithm. Argument identification and classification may be treated jointly or separately in the local scoring step. In the latter case, a pipeline of two sub-processes is typically applied, first scoring between “argument” and “no-argument” labels, and then scoring the particular argument labels. Because argument identification is closely related to syntax and argument classification is more a semantic issue, useful features for the two subtasks may be very different - that is, a good feature for addressing recognition may hurt classification and vice versa (Pradhan et al., 2005).
3. APPROACHES TO SEMANTIC ROLE LABELING

The third step in SRL is to apply a joint scoring (or global scoring) in order to combine the predictions of local scorers to produce a good structure of labeled arguments for the predicate. In this step, dependencies among several arguments of the same predicate can be exploited.

Some variations in the three-step architecture are found. Systems may bypass one of the steps, by doing only local scoring, or skipping directly to joint scoring. A fourth step may consist of fixing common errors or enforcing coherence in the final solution. This post-process usually consists of a set of hand-developed heuristic rules that are dependent on a particular architecture and corpus of application. An important consideration within this general SRL architecture is the combination of systems and input annotations.

Some SRL systems include some kind of combination to increase robustness, gain coverage, and reduce effects of parse errors. One may combine:

- the output of several independent SRL basic systems as in (Surdeanu et al., 2007), or
- several outputs from the same SRL system obtained by changing input annotations or other internal parameters, as discussed in (Toutanova et al., 2005).

The combination can be as simple as selecting the best among the set of complete candidate solutions, but usually consists of combining fragments of alternative solutions to construct the final output. Finally, the combination component may involve machine learning or not. The gain in performance when using the combination step is between two and three F1 points. However, a combination approach increases system complexity and penalizes efficiency.

Several exceptions to this described architecture for SRL can be found in the literature. One approach entails joint labeling of all predicates of the sentence, instead of proceeding one by one. This opens the possibility of exploiting dependencies among the different verbs in the sentence. However, the complexity may grow significantly, and results so far are inconclusive (Carreras et al., 2004; Surdeanu et al., 2007). Other promising approaches draw on dependency parsing rather than traditional phrase structure parsing (Johansson and Nugues, 2007), or combine parsing and SRL into a single step of semantic parsing (Musillo and Merlo, 2006).
3.1.4 Set of Features

As previously noted, devising the features with which to encode candidate arguments is crucial for obtaining good results in the SRL task. Given a verb and a candidate argument (a syntactic phrase) to be classified in the local scoring step, three types of features are typically used:

- features that characterize the candidate argument and its context;
- features that characterize the verb predicate and its context;
- features that capture the relation (either syntactic or semantic) between the candidate and the predicate.

Gildea and Jurafsky (2002b) presented a compact set of features across these three types, which has served as the core of most of the subsequent SRL work (see Section 3.2): (1) the phrase type, headword, and governing category of the constituent; (2) the lemma, voice, and subcategorization pattern of the verb; and (3) the left/right position of the constituent with respect to the verb, and the category path between them.

Extensions to these features have been proposed in various directions. Exploiting the ability of some machine learning algorithms to work with very large feature spaces, some authors have largely extended the representation of the constituent and its context, including among others: first and last words (and part-of-speeches) in the constituent, bag-of-words, n-grams of part of speech, and sequence of top syntactic elements in the constituent. Parent and sibling constituents in the tree may also be codified with all the previous structural and lexical features (Pradhan et al., 2005; Surdeanu et al., 2007).

Other authors have designed new features with specific linguistic motivations. For instance, Surdeanu et al. (2003) generalized the concept of headword with the content word feature. They also used named entity labels as features. Xue and Palmer (2004) presented the syntactic frame feature, which captures the overall sentence structure using the verb predicate and the constituent as pivots. All these features resulted in a significant increase in performance. Finally, regarding the relation between the constituent and the predicate, several variants of Gildea and Jurafsky’s syntactic path have been proposed in the literature (e.g., generalizations to avoid sparsity, and adaptations to partial parsing).
3. APPROACHES TO SEMANTIC ROLE LABELING

Also, some attempts have been made at characterizing the semantic relation between the predicate and the constituent. In (Erk, 2007; Zapirain et al., 2007), selectional preferences between predicate and headword of the constituent are explored to generate semantic compatibility features. Using conjunctions of several of the basic features is also common practice. This may be very relevant when the machine learning method used is linear in the space of features.

3.2 Probability estimation

The foundations for automatic semantic role labeling for FrameNet were laid in 2002 by Gildea and Jurafsky. Their system is based on statistical classifiers and is capable of assigning FrameNet roles to syntactically parsed sentences automatically. Later, the system was adapted to the task of labeling PropBank argument structures in (Gildea and Palmer, 2002).

Gildea and Jurafsky (2002b) propose a method to model global dependencies from the FrameNet corpus by including a probability distribution over multi-sets of semantic role labels given a predicate. The system first passed sentences through an automatic parser, extracted syntactic features from the parses, and estimated probabilities for semantic roles from the syntactic and lexical features. The distribution over label multi-sets is estimated using interpolation of relative frequency and a back-off distribution which assumes each argument label is present or absent independently of the other labels.

In spite of the pioneering nature of their research, they achieved quite impressive results: 82% accuracy in identifying the semantic role of pre-segmented constituents, 65% precision and 61% recall at the task of simultaneously segmenting constituents and identifying their semantic role, using posterior probability distributions for the classification task.

This statistical technique of labeling predicate argument operates on the output of the probabilistic parser reported in (Collins, 1999). It consists of two tasks: (1) identifying the parse tree constituents corresponding to arguments of each predicate encoded in PropBank; and (2) recognizing the role corresponding to each argument. For each task, different classifiers are trained.

For example, the result of the first classifier on the sentence illustrated in Figure 3.1 is the identification of the two NPs as arguments. The second classifier
3.2 Probability estimation

assigns the specific roles given the predicate “eat”.

Figure 3.1: Example of Parse Tree obtained with Collins parser, input for the semantic parser in (Gildea and Jurafsky, 2002b)

3.2.1 Feature set

Probabilities of a parse constituent belonging to a given semantic role are calculated using the following features:

phrase type feature (pt) indicates the syntactic type of the phrase expressing the semantic roles: examples include noun phrase (NP), verb phrase (VP), and full sentences (S). Phrase types were derived automatically from parse trees generated by the parser, as shown in Figure 3.1. The parse constituent spanning each set of words annotated as an argument was found, and the constituent’s nonterminal label was taken as the phrase type. As an example of how this feature is useful, in communication frames, the Speaker is likely to appear as a noun phrase, Topic as a prepositional phrase or noun phrase, and Medium as a prepositional phrase, as in:

[We]_{Speaker} talked [about the proposal]_{Topic} [over the phone]_{Medium}.

When no parse constituent was found with boundaries matching those of an argument during testing, the largest constituent beginning at the ar-
3. APPROACHES TO SEMANTIC ROLE LABELING

government’s left boundary and lying entirely within the element was used to calculate the features.

governing category (gov), has only two values, S and VP, corresponding to subjects and objects of verbs, respectively. This feature is restricted to apply only to NPs, as it was found to have little effect on other phrase types. As with phrase type, the feature was read from parse trees returned by the parser. Links from child to parent were followed up the parse tree from the constituent corresponding to a frame element until either an S or VP node is found, and assign the value of the feature according to whether this node is an S or VP. NP nodes found under S nodes are generally grammatical subjects, and NP nodes under VP nodes are generally objects. In most cases the S or VP node determining the value of this feature immediately dominates the NP node, but attachment errors by the parser or constructions such as conjunction of two NPs can cause intermediate nodes to be introduced. Searching for higher ancestor nodes makes the feature robust to such cases. Even given good parses, this feature is not perfect in discriminating grammatical functions, and in particular it confuses direct objects with adjunct NP such as temporal phrases. For example, town in the sentence He left town and yesterday in the sentence He left yesterday will both be assigned a governing category of VP. Direct and indirect objects both appear directly under the VP node. For example, in the sentence He gave me a new hose, me and a new hose are both assigned a governing category of VP.

parse tree path feature (path) is designed to capture the syntactic relation of a constituent to the predicate. It is defined as the path from the predicate through the parse tree to the constituent in question, represented as a string of parse tree nonterminals linked by symbols indicating upward or downward movement through the tree, as shown in Figure 3.2. Although the path is composed as a string of symbols, it can be also treated as an atomic value. The path includes, as the first element of the string, the part of speech of the predicate, and, as the last element, the phrase type or syntactic category of the sentence constituent marked as an argument.

position feature (p) simply indicates whether the constituent to be labeled
3.2 Probability estimation

Figure 3.2: Path Feature In this example, the path from the predicate *ate* to the argument NP *He* can be represented as VB ↑ VP ↑ S ↓ NP, with ↑ indicating upward movement in the parse tree and ↓ downward movement in the parsing tree (Gildea and Jurafsky, 2002b).

occurs before or after the predicate. This feature is highly correlated with grammatical function, since subjects will generally appear before a verb, and objects after. This feature may overcome the shortcomings of reading grammatical function from the parse tree, as well as errors in the parser output.

**voice feature** (v) distinguishes between active and passive verbs, and is important in predicting semantic roles because direct objects of active verbs correspond to subjects of passive verbs. An instance of a verb was considered passive if it is tagged as a past participle (e.g. *taken*), unless it occurs as a descendent verb phrase headed by any form of have (e.g. *has taken*) without an intervening verb phrase headed by any form of be (e.g. *has been taken*).

**head word** (h) is a lexical feature, and provides information about the semantic type of the role filler. Head words of noun phrases can be used to express selectional restrictions on the semantic types of role fillers. For example, in a *Communication frame*, noun phrases headed by *Bill, brother,* or *he* are more likely to be the *Speaker*, while those headed by *proposal, story,* or *question* are more likely to be the *Topic*. Head words of nodes in the parse tree are determined using the same deterministic set of head word
rules from Collins (1999) (usually nouns are head words for noun groups, verbs for verbal groups, prepositions are considered to be the head words of prepositional phrases and complementizers are considered to be heads for the groups they are in, meaning that infinitive verb phrases are always headed by to, and subordinate clauses such as in the sentence I’m sure that he came are headed by that).

3.2.2 Estimating probabilities

Given a predicate or target word \( t \) and a feature vector \( F \), containing the features presented above \( F = \{pt, gov, v, p, path, h\} \), the probability that a constituent fills the semantic role \( r \) is \( P(r|F,t) \). For example, suppose we only use the features \( gov \) (governing), \( v \) (voice) and \( pt \) (phrase type) to determine the role \( r \) of a constituent given the target word \( t \). The probabilities of the possible semantic roles are computed by counting the number of times each role appears in combination with these features, and dividing that by the total number of times the combination of features appears:

\[
P(r|gov,v,pt,t) = \frac{C(r,gov,v,pt,t)}{C(gov,v,pt,t)} \tag{3.1}
\]

where \( C \) represents the number of occurrences.

Using the training data, a distribution based on these probabilities can be computed. The model attempts to predict argument roles in new data, looking for the highest probability assignment of roles \( r_i \) to all constituents \( i \) in the sentence, given the set of features \( F_i = \{pt_i, path_i, pos_i, v_i, gov_i, h_i\} \) at each constituent in the parse tree, and the predicate \( t \):

\[
\text{argmax}_{r_1..n} P(r_{1..n}|F_{1..n}t) \tag{3.2}
\]

However, in many cases, a particular combination of feature may never be seen, or it may be seen only in a insignificantly law number of cases. The small number of training sentences for each target word and the large number of values that the head-word in particular can take (virtually any word of the language) contribute to the sparsity of the data. Therefore, they build the classifier by combining probabilities from distributions conditioned on a variety of subsets of the features. The distributions calculated were simply the empirical distributions
3.2 Probability estimation

Table 3.1: Example of probabilities computed by Gildea and Jurafsky (2002b) for the training set. The governing category feature is defined only for noun phrases.

| $P(r|pt, gov, t)$ | Count in training data |
|-------------------|------------------------|
| $P(r = AGT|pt = NP, gov = S, t = abduct)$ | 0.46 |
| $P(r = THM|pt = NP, gov = S, t = abduct)$ | 0.54 |
| $P(r = AGT|pt = PP, t = abduct)$ | 0.33 |
| $P(r = THM|pt = PP, t = abduct)$ | 0.33 |
| $P(r = MNR|pt = ADVP, t = abduct)$ | 1.00 |

from the training data, i.e. occurrences of each role and each set of conditioning events (subsets of features) were counted in a table, and probabilities were calculated by dividing the counts for each role by the total number of occurrences for each conditioning events. An example of the probabilities computed by Gildea and Jurafsky (2002b) for the features $pt$ and $gov$ are given in the table 3.1.

3.2.3 Combining probability distributions

Some distributions are more reliable than others. The lexical head-word statistics are valuable when data are available, but are particularly sparse due to the large number of possible head-words. In order to combine the strengths of the various distributions, Gildea and Jurafsky (2002b) merged them in different ways. The simplest combination method is linear interpolation, which simply averages the probabilities given by each of the distributions. For example:

$$P(r|t) = \lambda_1 P(r|v) + \lambda_2 P(r|p) + \lambda_3 P(r|v,p) + \lambda_4 P(r|p,t) + \lambda_5 P(r|v,t) + \lambda_6 P(r|v,p,t)$$  \hspace{1cm} (3.3)

where $\sum_i \lambda_i = 1$. In this example the features voice and position are combined to form one probability.

Another combination method used by Gildea and Jurafsky (2002b) is a “back-off” combination method, which consisted in creating a lattice over the distributions for various subsets of features, shown in figure 3.3.

The lattice is used to select a subset of the available distributions to combine. The less specific distributions were used only when no data were present for more
3. APPROACHES TO SEMANTIC ROLE LABELING

The lattice presented in figure 3.3 represents just one way of choosing subsets of features. The design of a feature lattice can be thought of as choosing a set of feature sets. Once the probability distributions of the lattice have been chosen, the graph structure of the lattice is determined by the subsumption relations among the sets of conditioning events. Given a set of N conditioning variables, there are $2^N$ possible subsets.

The particular lattice in figure 3.3 was chosen to represent some expected interaction between features (it is expected that *position* and *voice* interact, hence are always used together, while the *head-word* and the *phrase type* are relatively independent). The selected probabilities are combined using linear interpolation and geometric mean.

Gildea and Jurafsky experimented with 7 different combination methods: equal linear interpolation, EM linear interpolation, geometric mean, back off linear interpolation, back off geometric mean and assigning the most common role to a set of features. Their best performance on held out test data was achieved using a linear interpolation model in combination with the EM algorithm to calculate the weights of different feature vectors. Using basic features sets (combinations of phrase type, government, voice, position and target word), their final system performed at 80.4% accuracy.
3.3 Decision trees

In previous work using the PropBank corpus, Gildea and Palmer (2002) proposed a model predicting argument roles using the same statistical method as the one employed by Gildea and Jurafsky (2002b) for predicting semantic roles based on the FrameNet corpus. Statistical methods in general are hindered by the data sparsity problem. To achieve high accuracy and resolve the data sparsity problem, the method reported in (Gildea and Jurafsky, 2002b; Gildea and Palmer, 2002) employed a “back-off” solution based on a lattice that combines the model features. For practical reasons, this solution restricts the size of the feature sets. However, a larger set of features will determine a very complex lattice. Consequently, no new intuitions may be tested as no new features can be easily added to the model.

3.3.1 Feature set

To overcome this drawback, Surdeanu et al. (2003) found that inductive learning through decision trees allow to easily test large sets of features and study the impact of each feature on the augmented parser that outputs predicate-argument structures. Using the same sub-division of the labeling problem in two sub-problems, they used the C5 inductive decision tree learning algorithm (Quinlan, 1993) to implement both the classifier that identifies argument constituents and the classifier that labels arguments with their roles. Surdeanu et al. (2003) introduced a new set of features, after testing the set of features used in the work reported in (Gildea and Jurafsky, 2002b):

- **Content Word (cw)** - Lexicalized feature that selects an informative word from the constituent, different from the head word;

- **Part of Speech of Head Word (hPos)** - The part of speech tag of the head word;

- **Part of Speech of Content Word (cPos)** - The part of speech tag of the content word;

- **Named Entity class of content word (cNE)** - The class of the named entity that includes the content word.
3. APPROACHES TO SEMANTIC ROLE LABELING

- **Boolean Named Entity Flags** A feature set comprising:
  - `neOrganization`: set to 1 if an organization is recognized in the phrase
  - `neLocation`: set to 1 if a location is recognized in the phrase
  - `nePerson`: set to 1 if a person name is recognized in the phrase
  - `neMoney`: set to 1 if a currency expression is recognized in the phrase
  - `nePercent`: set to 1 if a percentage expression is recognized in the phrase
  - `neTime`: set to 1 if a time of day expression is recognized in the phrase
  - `neDate`: set to 1 if a date temporal expression is recognized in the phrase

- **Phrasal Verb Collocations** which comprises two features:
  - `pvcSum`: the frequency with which a verb is immediately followed by any preposition or particle.
  - `pvcMax`: the frequency with which a verb is followed by its predominant preposition or particle.

In developing the new set of features, the following observations on the data were considered important:

**Observation 1**

Because most of the predicate arguments are prepositional attachments (PP) or relative clauses (SBAR), often the head word (hw) feature is not in fact the most informative word in the phrase. Figure 3.4 illustrates three examples of this situation. In Figure 3.4(a), the head word of the PP phrase is the preposition `in`, but `June` is at least as informative as the head word. Similarly, in Figure 3.4(b), the relative clause is featured only by the relative pronoun `that`, whereas the verb `occurred` should also be taken into account. Figure 3.4(c) shows another example of an infinitive verb phrase, in which the head word is `to`, whereas the verb `declared` should also be considered. Based on these observations, Surdeanu et al. introduced the **content word** (cw), which adds a new lexicalization from
the argument constituent for better content representation. To select the content words they used the following heuristics:

**H1**: if phrase type is PP then select the rightmost child
Example: phrase = “in last June”, cw = “June”

**H2**: if phrase type is SBAR then select the leftmost sentence’ (S) head
Example: phrase = “that occurred yesterday”, cw = “occurred”

**H3**: if phrase type is VP then
    if there is a VP child then select the leftmost VP child, else select the head word
Example: phrase = “to be declared”, cw = “be declared” and recursively cw = “declared”

**H4**: if phrase type is ADVP then select the rightmost child not IN or TO
Example: phrase = “more than”, cw = “more”

**H5**: if phrase type is ADJP then select the rightmost adjective, verb, noun, or ADJP
Example: phrase = “61 years old”, cw = “old”

**H6**: for all other phrase types do select the head word

**Observation 2**

After implementing the set of features, the authors noticed that the \( hw \) feature was rarely used, because of data sparsity. Therefore they decided to add two new features, namely the parts of speech of the head word and the content word respectively.
3. APPROACHES TO SEMANTIC ROLE LABELING

Observation 3

Predicate arguments often contain names or other expressions identified by Named Entity (NE) recognizers, e.g. dates, prices. Thus Surdeanu et al. (2003) added the following features: (a) the named entity class of the content word (cNE); and (b) a set of NE features that can take only Boolean values. The cNE feature helps recognize the argument roles, e.g. ARGM-LOC and ARGM-TMP, when location or temporal expressions are identified. The Boolean NE flags provide information useful in processing complex nominals occurring in argument constituents.

Observation 4

Predicate argument structures are recognized accurately when both predicates and arguments are correctly identified. Often, predicates are lexicalized by phrasal verbs, e.g. put up, put off. To identify correctly the verb particle and capture it in the structure of predicates instead of the argument structure, they introduced two collocation features that measure the frequency with which verbs and succeeding prepositions co-occur in the corpus.

3.3.2 Feature Salience

The new features introduced by Surdeanu et al. (2003) increase the argument identification F-measure by 3.61%, and the role assignment accuracy with 4.29%. For the argument identification task, the head and content word features have a significant contribution for the task precision, whereas NE features contribute significantly to the task recall. For the role assignment task the best features from the Feature Set 2 are the content word features (cw and cPos) and the Boolean NE flags, which show that semantic information, even if minimal, is important for role classification.

The phrasal verb collocation features did not help for any of the tasks, but they were useful for boosting the decision trees. Decision tree learning provided by C5 (Quinlan, 2002) has built in support for boosting, and Surdeanu et al. (2003) used it and obtained improvements for both tasks. The best F-measure obtained for argument constituent identification was 88.98% in the fifth iteration (a 0.76% improvement). The best accuracy for role assignment was 83.74% in
The statistical model introduced in (Gildea and Jurafsky, 2002b) uses predicate lexical information at most levels in the probability lattice; hence, its scalability to unknown predicates is limited. In contrast, the decision tree approach of Surdeanu et al. (2003) uses predicate lexical information only for 5% of the branching decisions recorded when testing the role assignment task, and only for 0.01% of the branching decisions seen during the argument constituent identification evaluation.

3.4 Support Vector Machines (SVMs) approaches

In previous research, SVMs (Boser et al., 1992; Cortes and Vapnik, 1995) – a relatively new classification method – performed well on text classification tasks (Joachims, 1998), which is probably why many researchers have used them for semantic role classification. Training instances in SVM classification are represented as sparse feature vectors $\vec{x}_1 \ldots \vec{x}_n$ in a high dimensional space $\chi \subseteq \mathbb{R}^d$. Suppose training instances belong either to positive or negative class as follows:

$$(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \text{ with } x_i \in \chi, \ y_i \in \{+1, -1\} \quad (3.4)$$

$\vec{x}_i$ is a $d$ dimensional feature vector of the $i$-th sample, $y_i$ is a scalar value that specifies the class (positive or negative) of $i$-th data. Classification can be described as building the function $f : \chi \rightarrow \{ \pm 1 \}$ by going through a learning process, under the assumption that new examples were generated from the same unknown probability distribution $P(\vec{x}_i, y_i)$.

A SVM classifier tries to separate positive from negative examples by finding a hyperplane\(^5\) that divides the training data in two groups:

$$w \cdot x + b = 0 \text{ where } w \in \mathbb{R}^n, \ b \in \mathbb{R}. \quad (3.5)$$

SVMs try to optimize parameters $w$ and $b$ to find the optimal solution. Optimal means finding the hyperplane separating training examples with the maximal margin. Margin can be defined informally by the distance between the hyperplane and the boundaries in which it can move without any misclassification.

\(^5\)a $(d - 1)$-dimensional hyperplane for data points represented in a $d$-dimensional space.
3. APPROACHES TO SEMANTIC ROLE LABELING

A problem with using SVMs for the SRL task, is that SVMs are binary classifiers, i.e. they can only differentiate between two classes. Two approaches are possible:

• Pairwise approach - A separate binary classifier is trained for each of the class pairs and their outputs are combined to predict the classes. The total number of classifiers required for this approach is $\frac{N(N-1)}{2}$, where $N$ is the number of roles to be classified.

• One versus all (OVA) approach - $N$ classifiers are trained for a class problem. Each classifier can discriminate between a particular class and the set of all other classes.

Pradhan et al. (2005) used One-vs.-all formalism to train SVM for Semantic Role Labeling. They divided the labeling task in three sub-tasks, training different classifiers for each task:

**Argument Identification** : This is the process of identifying parsed constituents in the sentence that represent semantic arguments of a given predicate.

**Argument Classification** : Given constituents known to represent arguments of a predicate, assign the appropriate argument labels to them.

**Argument Identification and Classification** : A combination of the above two tasks.

Another problem is that kernel-based systems such as SVMs, exhibit classification speeds which are substantially lower than those of other machine learning algorithms (Kudo, 2003). Pradhan et al. (2005) divided the training process into two stages. First they filter out the nodes that have a very high probability of being NULL. Once the sentence has been parsed using a parser, each node in the parse tree can be classified as either one that represents a semantic argument (i.e., a NON-NULL node) or one that does not represent any semantic argument (i.e., a NULL node). The NON-NULL nodes can then be further classified into the set of argument labels. For example, following Pradhan et al. (2005), in the sentence:

(3)  **He talked for about 20 minutes.**

68
3.4 Support Vector Machines (SVMs) approaches

the PP that encompasses for about 20 minutes is a NULL node because it does not correspond to a semantic role (due to the preposition for). The node NP that encompasses about 20 minutes is a NON-NULL node, since it does correspond to a semantic role, the ARGM-TMP$^6$.

A binary NULL vs NON-NULL classifier is trained on the entire dataset. The remaining training data is used to train OVA classifiers, one of which is the NULL–NON-NULL classifier. With this strategy only one classifier (NULL vs NON-NULL) has to be trained on all of the data. The remaining OVA classifiers are trained on the nodes passed by the filter (approximately 20% of the total), resulting in a considerable savings in training time.

The SVM model was built using TinySVM2 along with YamCha3 (Kudo and Matsumoto, 2002; Kudo, 2003) as the SVM training and test software. The system uses a polynomial kernel$^7$ with degree 2; the cost per unit violation of the margin $C = 1$; and tolerance of the termination criterion $\epsilon = 0.001$.

3.4.1 Feature Set

The model of Pradhan et al. is a state-of-the-art model, incorporating a large set of structural and lexical features.

Pradhan et al. (2005) used the features proposed by Gildea and Jurafsky (2002b), the named entities in constituents and head word part of speech introduced by Surdeanu et al. (2003), and some novel features:

$^6$In our approach (presented in Chapter 4), the whole prepositional group for about 20 minutes is a NON-NULL role, while the noun group without the preposition for is a NULL role. We adopted this opposite approach since we consider prepositions very useful for determining the semantic roles type.

$^7$In their basic form, SVMs learn linear threshold function. Nevertheless, vectors $x_i$ may be mapped into higher dimensional spaces by replacing every dot product $w \cdot x$ with a non-linear kernel function. Some kernel functions have been found to work well for a wide variety of applications: polynomial, radial basis function (RBF) or sigmoid functions.

$^8$To allow some flexibility in separating the categories, SVM models have a cost parameter, $C$, that controls the trade-off between allowing training errors and forcing rigid margins. This creates a soft margin that permits some misclassification. Increasing the value of $C$ increases the cost of misclassifying points and forces the creation of a more accurate model, which may not generalize well.

$^9$A kernel parameter used by TinySVM2 to specify the number of iterations to perform in order to have as few misclassification as possible, yet avoiding over-fitting.
3. APPROACHES TO SEMANTIC ROLE LABELING

Verb Clustering: Since they used limited training data, any real world test set will contain predicates that have not been seen in training. In these cases, to benefit from some information about the predicate, they use predicate cluster as a feature. The verbs were clustered into 64 classes using the probabilistic co-occurrence model of Hofmann and Puzicha (1998). The clustering algorithm uses a database of verb-direct-object relations extracted by Lin (1998). The obtained verb class of the current predicate was used as a feature.

Partial Path: For the argument identification task, path is the most salient feature. However, it is also the most data sparse feature. To overcome this problem, they generalize the path by adding a new feature that contains only the part of the path from the constituent to the lowest common ancestor of the predicate and the constituent, called “Partial Path”.

Verb Sense Information: The arguments that a predicate can take depend on the word sense of the predicate. Each predicate tagged in the PropBank corpus is assigned a separate set of arguments depending on the sense in which it is used. For instance, depending on the sense of the predicate talk, either ARG1 or ARG2 can identify the Hearer. Each sense of a predicate was considered to belong to a different class.

Head Word of Prepositional Phrases: Many adjunctive arguments, such as temporal and locative circumstantial phrases, occur as prepositional phrases in a sentence, and it is often the case that the head words of those phrases, which are always prepositions, are not very discriminative, eg., in the city, in a few minutes, both share the same head word in and neither contain a named entity, but the former is ARGM-LOC, whereas the latter is ARGM-TMP. Therefore, Pradhan et al. (2005) replaced the head word of a prepositional phrase, with that of the first noun phrase inside the prepositional phrase. The preposition information was retained by appending it to the phrase type, eg., “PP-in” instead of “PP” for the examples above.

First and Last Word/POS in Constituent: Some arguments tend to contain discriminative first and last words.
3.4 Support Vector Machines (SVMs) approaches

**Ordinal constituent position**: In order to avoid false positives of the type where constituents far away from the predicate are spuriously identified as arguments, this feature included the concatenation of the constituent type and its ordinal position with regard to the predicate.

**Constituent relative features**: These are nine features representing the phrase type, head word and head word part of speech of the parent, and left and right siblings of the constituent in focus. These were added on the intuition that encoding the tree context this way might add robustness and improve generalization.

**Temporal cue words**: There are several temporal cue words that are not captured by the named entity tagger and were considered for addition as a binary feature indicating their presence.

**Dynamic class context**: In the task of argument classification, these are dynamic features that represent the classifications of the previous two nodes belonging to the same tree as the node being classified.

3.4.2 Feature Salience

Pradhan et al. (2005) tested the effect each feature has on the argument classification and the argument identification tasks. Addition of named entities improves the F1 score for adjunctive arguments ARGM-LOC from 59% to 68% and ARGM-TMP from 78.8% to 83.4%. But, since these arguments are small in number compared to the core arguments, the overall accuracy does not show a significant improvement. Adding this feature to the NULL vs NON-NULL classifier degraded its performance.

Replacing the head word and the head word POS for prepositional phrases by the head word of the noun phrase inside it had a positive impact on the overall results.

Two other ways of generalizing the head word were considered: i) adding the head word cluster as a feature, and ii) replacing the head word with a named entity if it belonged to any of the seven named entities mentioned earlier. Neither method showed any improvement. Nor did generalizing the path feature by i)
3. APPROACHES TO SEMANTIC ROLE LABELING

compressing sequences of identical labels, or ii) removing the direction in the path.

In order to improve the performance of their statistical argument tagger, Pradhan et al. (2005) used the fact that a predicate is likely to instantiate a certain set of arguments (extracted from the annotation examples found in the PropBank corpus). Using some additional constraints: i) argument ordering information is retained, and ii) the predicate is considered as an argument and is part of the sequence, they trained a trigram language model on the argument sequences to estimate the probability of argument sets not seen in the training data.

The performance of their system is the state-of-the-art performance in semantic role labeling, having, on hand-annotated data, precision 88.9%, recall 84.6%, F-measure 86.7% for both argument and adjunct types classification, and precision 90.5%, recall 87.4% and F-measure 88.9% for the task of only classifying arguments, given a predicate and its frame.

3.5 Memory Based Learning (MBL)

Memory based learning can be described as reasoning on the basis of similarity of new situations to earlier encountered situations (Daelemans et al., 2003). The learning component of a MBL system stores the examples in the memory. The other component of a MBL system, the performance component, is similarity-based and performs the actual classification.

During training, training instances are loaded into memory. An instance consists of a vector containing feature-value pairs and a class assignment. During classification, unseen examples are compared to instances in the training data. This comparison is done using a distance metric $\Delta(X, Y)$. The class assignment is based on the $k$-nearest neighbors algorithm\textsuperscript{10}: the most common class amongst

\textsuperscript{10}In pattern recognition, the $k$-nearest neighbors algorithm ($k$-NN) is a method for classifying objects based on closest training examples in the feature space. $k$-NN is a type of instance-based learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its $k$ nearest neighbors ($k$ is a positive integer, typically small).
3.5 Memory Based Learning (MBL)

the $k$ most similar training instances is chosen. In case of a tie among categories, a tie breaking resolution method is used.

Different distance metrics can be used in MBL. The most common of which is the overlap metric:

$$\Delta(X,Y) = \sum_{i=1}^{n} w_i \delta(x_i, y_i)$$  \hspace{1cm} (3.6)

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases} \hspace{1cm} (3.7)$$

where $\Delta(X,Y)$ is the distance between instances $X$ and $Y$, represented by $n$ features, $\delta$ is the distance per feature, and $w_i$ is a weight parameter.

A $k$-NN algorithm with this metric is called IB1. In the overlap metric, all features have the same weight. To improve performance, domain knowledge can be used to assign different weights to different features. A useful tool for measuring feature relevance that is especially beneficial for NLP tasks, is information gain (IG) (Daelemans et al., 2003).

Information Gain (IG) looks at each feature and measures how much information it contributes to the knowledge needed to predict the correct class label. The most common way of measuring the IG of a feature $i$ is to compute the difference in uncertainty (i.e. entropy) between situations without and with knowledge of the value of that feature:

$$w_i = H(C) - \sum_{v \in V_i} P(v) \times H(C|v)$$  \hspace{1cm} (3.8)

where $C$ is the set of class labels, $V_i$ is the set of values $v$ for the feature $i$, $H(C)$ is the entropy of the class labels, $P(v)$ is the probability of the feature $i$ having the value $v$, and $H(C|v)$ is the entropy of the class labels given the value $v$. Entropy measures the amount of uncertainty of a variable, and is defined as:

$$H(C) = - \sum_{c \in C} p(c) \log_2 p(c)$$  \hspace{1cm} (3.9)
3. APPROACHES TO SEMANTIC ROLE LABELING

Morante et al. (2008) use IB1 classifier implemented in the TiMBL package (Daelemans et al., 2003) to process semantic dependencies for English. Their system handles semantic role labeling in three steps: predicate identification, semantic dependency classification, and combination of classifiers that predicted the semantic classification.

For the first task, the identification of the predicates, the IB1 algorithm was parameterized by using overlap as the similarity metric, information gain for feature weighting, 7 k-nearest neighbors, and weighting the class vote of neighbors as a function of their inverse linear distance. The instance represents all nouns and verbs in the corpus. The features used for the predicate identification task are: word form, lemma, part of speech, the three last letters of the word, and the lemma and part of speech of the five previous and five next words.

For the semantic dependency classification task, Morante et al. (2008) uses three groups of multi-class classifiers to predict, in one step, if there is a dependency between a word and a predicate, and the type of the dependency, i.e. the semantic role. The first group of classifiers consists of two IB1 classifiers, one for noun and one for verbal predicates. The instances represent a predicate-word combination, where the predicates are the ones identified in the previous task, and the candidate words are filtered to eliminate the determiners and other functional words that cannot bear semantic roles. The classifiers in this group use 11 k-nearest neighbors and the following features: word form, part of speech, dependency type, word form of the two previous and two next words, chain of POS types between the word and the predicate, distance between the word and the predicate, a binary feature indicating if the word depends on the predicate, six chains of part of speech tags between the word and its three previous, respectively three next predicates in relation to the current predicate.

The second group of classifiers consists also of two IB1 classifiers, one for noun and one for verb predicates, with the difference that the classifiers used 7 k-nearest neighbor version of the algorithm. The instances represent predicate-verb combinations, but only the ones that were marked as having a semantic role by the classifiers in the first group. The features used by these classifiers are: word form, chain of lemmas of the syntactic siblings, chain of lemmas of the syntactic children, dependency type, word form of the two previous and the two next words, part of speech and type of dependency predicted by the classifier.
in the first group, lemma of the syntactic head, chain of dependency types and chain of lemmas of the syntactic children, chain of part of speech types that exists in the linear representation of the sentence between the word and the predicate, distance between the word and the predicate.

The third group of classifiers is similar with the second one as to the features used and the classifier types, the only difference is only one classifier is trained for both predicate types (noun or verb).

The last task consists in a simple combination of the predictions of the classifiers in the three groups: if group 2 and 3 agree in classifying a semantic dependency, their solution is chosen, else the solution of the classifier in group 1 is chosen, since this group has a higher accuracy. The overall F-measure of their system is 69.75%.

3.6 Conclusions

Assignment of semantic roles is an important part of language understanding, and has been attacked by many computational systems. Traditional parsing and understanding systems, including implementations of unification-based grammars such as HPSG (Pollard and Sag, 1994), rely on hand-developed grammars which must anticipate each way in which semantic roles may be realized syntactically. Writing such grammars is time-consuming, and typically such systems have limited coverage. Data-driven techniques have recently been applied to template-based semantic interpretation in limited domains by “shallow” systems that avoid complex feature structures, and often perform only shallow syntactic analysis. They show promise for a more sophisticated approach to generalize beyond the relatively small number of frames considered in the tasks.

This chapter has presented the characteristics that a semantic role labeling system must have, starting with the types of vector spaces used in textual machine learning, the “standard” architecture of a data-driven semantic role labeling system, and an inventory of features used in the literature. The different machine learning strategies used (probability estimation, decision trees, support vector machine and memory-based learning) are also presented.

An important drawback of the three presented systems is that they don’t treat nominal predicates, being only built for verbal predicates. Furthermore, they only
consider one predicate per sentence, even if this is not always the case. For example, in the sentence  

The awarding of the Nobel Prize to President Obama was largely debated, 

we have two predicational words, the awarding, having as ARG1 of the Nobel Prize, and debated, having two arguments, an ARG1 The awarding of the Nobel Prize to President Obama and ARGM-MNR largely.

When comparatively analyzing the presented methods, the best results were reported with SVMs, the other methods having however close results. However, SVMs training time can be considerable long, especially on large data sets. Considering, however, that Pradhan et al. (2005) highly optimized the feature set, and that by applying pre- and post-processing filters the results of the other methods could be improved, we envisaged testing whether other machine learning algorithms can perform similar to SVMs, or even better.
Development of a Semantic Role Labeling System

Recognizing and labeling semantic arguments is a key task for answering “Who”, “When”, “What”, “Where”, “Why”, etc. questions in Information Extraction, Question Answering, Summarization, and, in general, in all NLP tasks in which some kind of semantic interpretation is needed. The importance of the task is confirmed by the emergence of international competitions concerned with the recognition of semantic roles (mainly for the English language), based on PropBank predicate-argument structures (Shared Tasks of CoNLL 2005-2009\(^1\)) or FrameNet semantic frames (SemEval 2007\(^2\)). Given a sentence, the task consists of analyzing the scene expressed by some predicational word of the sentence. In particular, for each predicational verb or noun (referred as target word), all the constituents in the sentence which fill a semantic role of the target word have to be recognized and labeled accordingly. This problem has been referred already in this thesis as Semantic Role Labeling (SRL) and the most common methods to perform SRL have been presented in Chapter 3.

One of the main ambitions of this thesis is to develop a Semantic Role Labeling System to be incorporated in the applications developed by the Natural Language Processing Group at the Faculty of Computer Science. In order to address this task, we started by building a rule-based SRL system. However, due to the large

---

\(^1\)CoNLL web page: http://ifarm.nl/signll/conll/
\(^2\)SemEval 2007 web page: http://framenet.icsi.berkeley.edu/semeval/FSSE.html
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

size and the variety of the training data, the performances of the rule-based SRL system were limited; therefore, a statistical approach was considered. In order to identify which of the machine learning techniques (among those investigated in Chapter 3) is best suited for the SRL task, we have tested several techniques for different languages, using the algorithms implemented in the Weka\textsuperscript{3} toolkit. The final system developed into a “platform” for creating supervised Semantic Role Labeling systems starting with an annotated corpus. The platform trains several classifiers, chooses the ones with the greatest performance and returns a Semantic Role Labeling System (a sequence of trained models to run on new, unannotated data).

Both systems (the rule-based and the learning-based) receive as input plain, unannotated text, add syntactic annotation at a pre-processing step, and output an XML file where syntactic constituents are annotated with their corresponding semantic roles, if any. The rule-based system was developed during the CoNLL-2008 competition for English, while the statistical system was developed for several languages (English, German, Czech, Japanese, Chinese), using the training data from the CoNLL-2009 competition.

4.1 RuleSRL: A Rule-based Semantic Role Labeling System

4.1.1 Input data

Our first attempt to create a SRL system for English involved developing a rule-based system (referred to as RuleSRL from now on) that attached a semantic role to each syntactic constituent. Since the input of RuleSRL is a raw text, two pre-processing steps were needed: adding part of speech information and adding syntactic dependencies. The part of speech information is needed in order to decide which words are going to be considered possible predicational target words\textsuperscript{4}, while the syntactic dependencies, used in a form or another from the first

\textsuperscript{3}Weka web page: http://www.cs.waikato.ac.nz/ml/index.html

\textsuperscript{4}Although, as presented in section 2.1.4, adjectives can also have predicational behavior, only verbs and nouns were considered for the development of our SRL systems, since they are more frequent, and large annotated resources exists only for these two predication types.
4.1 RuleSRL: A Rule-based Semantic Role Labeling System

SRL systems\(^5\), are proven to be important for the SRL systems (Punyakanok et al., 2008).

The part of speech information for English is obtained using the Stanford Parser (Klein and Manning, 2003a, b). The Stanford Parser is a probabilistic parser that works out the grammatical structure of sentences, adding part of speech and chunking information to plain text. It is a Java implementation of a probabilistic natural language parser, both highly optimized, based on Probabilistic Context-Free Grammars (PCFG)\(^6\).

For the sentence:

\[1\] The economy’s temperature will be taken from several vantage points this week, with readings on trade, output, housing and inflation.

the part of speech annotation obtained with the Stanford Parser is:

\[2\] The/DT economy/NN ’s/POS temperature/NN will/MD be/VB taken/VBN from/IN several/JJ vantage/NN points/NNS this/DT week/NN ,/, with/IN readings/NNS on/IN trade/NN ,/, output/NN ,/, housing/NN and/CC inflation/NN ./.

where each word is followed by its part of speech tag.

Beside part of speech, the input was processed also for dependency structure. Although the Stanford Parser is able to add also dependency relations, we preferred to use the MaltParser (Nivre, 2003), due to its higher accuracy and because it outputs a dependency tree, which is not always the case for the Stanford Parser. MaltParser is a system for data-driven dependency parsing, which can be used to induce a parsing model from treebank data or to parse new data using an induced model\(^7\). While a traditional parser-generator constructs a parser, given a grammar, a data-driven parser-generator constructs a parser, given a treebank. MaltParser is an implementation of inductive dependency parsing, where

\(^5\)Gildea and Jurafsky (2002b) used the head-word, which, although defined using the information available in the syntactic parse tree, represents a form of simplified dependency.

\(^6\)The original version of this parser was mainly written by Dan Klein, with support code and linguistic grammar development by Christopher Manning.

\(^7\)MaltParser is developed by Johan Hall, Jens Nilsson and Joakim Nivre at Växjö University and Uppsala University, Sweden.
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

the syntactic analysis of a sentence amounts to the derivation of a dependency structure, and where inductive machine learning is used to guide the parser at nondeterministic choice points (Nivre, 2006). The parsing methodology is based on three essential components:

1. Deterministic parsing algorithms for building labeled dependency graphs (Nivre, 2003; Yamada and Matsumoto, 2003);

2. History-based models for predicting the next parser action at nondeterministic choice points (Black et al., 1992; Collins, 1999; Magerman, 1995);

3. Discriminative learning to map histories to parser actions (Hall et al., 2006; Kudo and Matsumoto, 2002; Yamada and Matsumoto, 2003).

The model we used for English was trained using the Wall Street Journal section of the Penn Treebank corpus, annotated with dependency relations, available within the CoNLL-2008 competition as training set.

<table>
<thead>
<tr>
<th>Field number</th>
<th>Field name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>Token counter, starting at 1 for each new sentence.</td>
</tr>
<tr>
<td>2</td>
<td>FORM</td>
<td>Word form or punctuation symbol.</td>
</tr>
<tr>
<td>3</td>
<td>LEMMA</td>
<td>Lemma or stem (depending on particular data set) of word form, or an underscore, if not available.</td>
</tr>
<tr>
<td>4</td>
<td>POS</td>
<td>Part-of-speech tag, where the tagset depends on the language, or identical to the coarse-grained part-of-speech tag, if not available.</td>
</tr>
<tr>
<td>5</td>
<td>HEAD</td>
<td>Head of the current token, which is either a value of ID (meaning that the word with the ID (n) is the head-word of the current token) or zero (’0’, if the head word is the ROOT fictive node).</td>
</tr>
<tr>
<td>6</td>
<td>DEPREL</td>
<td>Dependency relation to the HEAD. The set of dependency relations depends on the particular language.</td>
</tr>
</tbody>
</table>
After the pre-processing step, the English sentence contains syntactic information, the main type of data needed as input by RuleSRL. The information was stored in a column-style format, since this was the standard used within the CONLL competitions (also the format of the training data). The columns of the input file for the RuleSRL system are presented in table 4.1.

![Figure 4.1: Example of Dependency Relations annotated using the Malt-Parse](image)

The ID of the words is initialized to 1 for each new sentence. The LEMMA is obtained by a simple lemmatizer which maps the current word against a list of English inflected words, the POS is obtained with the Stanford Parser, and the HEAD and DEPREL are obtained with the MaltParser.

![Figure 4.2: Example of Dependency Tree for a sentence annotated using the MaltParser](image)

An example of annotated sentence is given in Figure 4.1 for the sentence:
The economy’s temperature will be taken from several vantage points this week, with readings on trade, output, housing and inflation.

The annotation behind the figure is presented in table 4.2, in a column-like format. Figure 4.2 represents graphically the dependency tree for the sentence (3).

**Table 4.2: Example of Syntactic and Dependency parsed sentence**

<table>
<thead>
<tr>
<th>ID</th>
<th>Form</th>
<th>Lemma</th>
<th>POS</th>
<th>HeadID</th>
<th>Deprel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The</td>
<td>the</td>
<td>DT</td>
<td>2</td>
<td>NMOD</td>
</tr>
<tr>
<td>2</td>
<td>economy</td>
<td>economy</td>
<td>NN</td>
<td>4</td>
<td>NMOD</td>
</tr>
<tr>
<td>3</td>
<td>’s</td>
<td>’s</td>
<td>POS</td>
<td>2</td>
<td>SUFFIX</td>
</tr>
<tr>
<td>4</td>
<td>temperature</td>
<td>temperature</td>
<td>NN</td>
<td>5</td>
<td>SBJ</td>
</tr>
<tr>
<td>5</td>
<td>will</td>
<td>will</td>
<td>MD</td>
<td>0</td>
<td>ROOT</td>
</tr>
<tr>
<td>6</td>
<td>be</td>
<td>be</td>
<td>VB</td>
<td>5</td>
<td>VC</td>
</tr>
<tr>
<td>7</td>
<td>taken</td>
<td>take</td>
<td>VBN</td>
<td>6</td>
<td>VC</td>
</tr>
<tr>
<td>8</td>
<td>from</td>
<td>from</td>
<td>IN</td>
<td>7</td>
<td>CLR</td>
</tr>
<tr>
<td>9</td>
<td>several</td>
<td>several</td>
<td>JJ</td>
<td>11</td>
<td>NMOD</td>
</tr>
<tr>
<td>10</td>
<td>vantage</td>
<td>vantage</td>
<td>NN</td>
<td>11</td>
<td>NMOD</td>
</tr>
<tr>
<td>11</td>
<td>points</td>
<td>point</td>
<td>NNS</td>
<td>8</td>
<td>PMOD</td>
</tr>
<tr>
<td>12</td>
<td>this</td>
<td>this</td>
<td>DT</td>
<td>13</td>
<td>NMOD</td>
</tr>
<tr>
<td>13</td>
<td>week</td>
<td>week</td>
<td>NN</td>
<td>7</td>
<td>TMP</td>
</tr>
<tr>
<td>14</td>
<td>,</td>
<td>,</td>
<td>,</td>
<td>7</td>
<td>P</td>
</tr>
<tr>
<td>15</td>
<td>with</td>
<td>with</td>
<td>IN</td>
<td>7</td>
<td>ADV</td>
</tr>
<tr>
<td>16</td>
<td>readings</td>
<td>reading</td>
<td>NNS</td>
<td>15</td>
<td>PMOD</td>
</tr>
<tr>
<td>17</td>
<td>on</td>
<td>on</td>
<td>IN</td>
<td>16</td>
<td>NMOD</td>
</tr>
<tr>
<td>18</td>
<td>trade</td>
<td>trade</td>
<td>NN</td>
<td>17</td>
<td>PMOD</td>
</tr>
<tr>
<td>19</td>
<td>,</td>
<td>,</td>
<td>,</td>
<td>18</td>
<td>P</td>
</tr>
<tr>
<td>20</td>
<td>output</td>
<td>output</td>
<td>NN</td>
<td>18</td>
<td>COORD</td>
</tr>
<tr>
<td>21</td>
<td>,</td>
<td>,</td>
<td>,</td>
<td>20</td>
<td>P</td>
</tr>
<tr>
<td>22</td>
<td>housing</td>
<td>housing</td>
<td>NN</td>
<td>20</td>
<td>COORD</td>
</tr>
<tr>
<td>23</td>
<td>and</td>
<td>and</td>
<td>CC</td>
<td>22</td>
<td>COORD</td>
</tr>
<tr>
<td>24</td>
<td>inflation</td>
<td>inflation</td>
<td>NN</td>
<td>23</td>
<td>CONJ</td>
</tr>
<tr>
<td>25</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>5</td>
<td>P</td>
</tr>
</tbody>
</table>
Thus, if looking at the figure and at the table 4.2, we see for instance, that the word The, having the ID 1, has as head the word with the ID 2, economy. One important observation is that the annotation of the verb group is sequential, thus the target word taken is a child of be, who is a child of will, although they form together the verbal group will be taken. This annotation must be corrected before the semantic role assignment phase, otherwise the noun group the economy’s temperature will not be recognized as a semantic role of the verb taken, since it is not a direct child of it. At the same time, the noun group the economy’s temperature will not be a semantic role of the verb will either, since auxiliaries take no semantic roles. This situation will be discussed in Section 4.1.2.1.

4.1.2 The System’s Architecture

The architecture of the rule-based system for Semantic Role Labeling is presented in figure 4.3. The system is composed out of three main modules:

**Predicate Identification** – this module takes the syntactic analyzed sentence and decides which of its verbs and nouns are predicational, thus for which ones semantic roles need to be identified;

**Predicate Sense Identification** – once the predicates for a sentence are marked, each predicate need to be disambiguated since, for a given predicate, different sense may demand different types of semantic roles;

**Semantic Roles Identification** – identify the semantic roles for each of the syntactic dependents of the selected predicates, and establish what kind of semantic roles it is (what label it carries).

4.1.2.1 Predicate Identification

Our rule-based semantic role labeling system uses the PropBank annotation of semantic roles. Since predicational words are not just verbs, beside PropBank (Palmer et al., 2005) for the verbal frames, NomBank (Meyers et al., 2004a) is also
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

Figure 4.3: Rule-based SRL System Architecture for English

used for nouns. Starting with the syntactic annotation presented in the previous section, with marked dependency relations, and using as resources PropBank and NomBank, the system first tries to identify the words in the sentence that can be semantic predicates, and for which semantic roles need to be found and annotated. This module relies mainly on the external resources, thus the verbs that are in PropBank (have semantic frame annotation) are likely to be semantic predicates, those which aren’t, are not predicational verbs, thus cannot have semantic arguments. For example, the verb \textit{to be} has no annotation in PropBank, since it is a state and not an action, predicational, verb. Similarly, the NomBank is used to sort nouns that can behave as predicates from those that cannot have semantic arguments. The output of this module is the input file, where each verb
4.1 RuleSRL: A Rule-based Semantic Role Labeling System

or noun that behaves as a predicate is annotated.

Another important factor in deciding if the verb/noun can play the target predicate role is checking if it is head for any syntactic constituents. Even if the verb or noun has annotations is the external resources, and it may behave as a predicate, if it has no dependents, there is no point in annotating it as predicative, since there will be no semantic frame to be annotated for that target word\(^8\). The only exception that we allow is in the case of verbs with auxiliaries or modal verbs, since sometimes the arguments are linked to the auxiliary verb, instead of the main verb. For example, if we look back at the Table 4.2, one can notice that the syntactic dependency annotation links the noun temperature as subordinate of the auxiliary verb will, although will is only part of the verbal group will be taken and the main verb taken is in fact the head of temperature.

After the predicates from the input sentence are identified, the next two modules are successively applied for all the predicates in the sentence, in order to identify for all of them all and only their arguments. For example, for the sentence:

(4) The assignment of semantic roles depends on the number of predicates.

two predicates are identified by the Predicate Identification module (assignment and depend). The next two modules will be applied two times, once for the identification of the sense and semantic arguments of the assignment predicate, and once for the depend predicate. The output of RuleSRL will therefore provide the two annotations:

(5) [The assignment of semantic roles]\(_{ARG0}\) \([\text{depends}]_{TARGET}\) \([\text{on the number of predicates}]_{ARG1}\).

(6) [The assignment]\(_{TARGET}\) \([\text{of semantic roles}]_{ARG1}\) depends on the number of predicates.

---

\(^8\)The PropBank annotation allows for uninstantiated roles to be annotated. For example, in the sentence I want [*] and like to eat ice-cream, the ARG1 of the verb want is to eat ice cream, but since it is the same as ARG1 of the verb like, it is present in the sentence only once. However, RuleSRL has rules only for the instantiated (lexically present) roles.
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

4.1.2.2 Predicate Sense Identification

Once the predicate is identified, the next step involved determining which sense the predicate has, according to the PropBank / NomBank annotation, in order to select the types of semantic roles its sense allow for. An example of the PropBank annotation of a predicate’s senses, a small part of the PropBank annotation for the verb *take* is presented in figure 4.4. As it can be seen from figure 4.4, the annotations for a target word (either verb or noun), in this case the verb *take*, are grouped with regard to its lemma in different *predicate* tags. The *lemma* attribute of the *predicate* tag is considered either strictly the lemma of the lexical verb, or the lemma followed by an adverbial particle, if the verb allows for any (i.e. the verb is a phrasal verb). For example, for the verb *take*, 9 predicate tags are annotated, with the lemmas: *take*, *take away*, *take in*, *take off*, *take on*, *take out*, *take over*, *take up* and *take aback*\(^9\). The first predicate, corresponding to the lexical verb *take* with no particles, has nine senses annotated in PropBank (two of which are presented in Figure 4.4) as role sets. Thus, for the verb *take*, the nine senses of the lexical lemma without any particles are: (1) take, acquire, come to have, (2) tolerate, (3) cause (to be), (4) understand to be, (5) fixed phrase: take place, (6) write-off, acknowledge a (financial) loss, (7) need, (8) take by surprise, (9) become fond of.

The sense annotation in PropBank and NomBank is similar to some extend to the sense annotation from WordNet, with the observation that the classification in sense classes (role sets in PropBank’s terminology) is centered less on the different meanings of the predicational word, and more on the difference between the sets of semantic roles that two senses may have. The senses and role sets in PropBank for a particular predicate are usually subsumed by WordNet senses, since the latter has a finer sense distinction. For instance, the *predicate* tag for the verb *take* with the lemma *take* (the role set where *take* has no phrasal form) has in PropBank 9 role sets (see Appendix A), and in WordNet 42 synsets (see Appendix B). As an example of the correspondence between the two resources, let’s consider the first role set in PropBank for the *predicate* tag with the lemma *take*: *take, acquire, come to have* with their annotated examples. For each

\(^9\)For formatting reasons, figure 4.4 presents only two of the nine *predicate* tags that exists in the PropBank frame for the verb *take*. The entire PropBank entry for the verb *take* is given in Appendix A.
4.1 RuleSRL: A Rule-based Semantic Role Labeling System

Figure 4.4: Example of PropBank Annotation for the verb take

example, the synsets from WordNet that have similar meaning are listed:

- take office “President Carlos Menem took office July 8.”

Possible WordNet correspondences: fill, take, occupy (assume, as of positions or roles) “She took the job as director of development”; “he occupies the position of manager”; “the young prince will soon occupy the
throne”;

WN: assume, take, strike, take up (occupy or take on) “He assumes the lotus position”; “She took her seat on the stage”; “We took our seats in the orchestra”; “She took up her position behind the tree”; “strike a pose”

WN: take (take by force) “Hitler took the Baltic Republics”; “The army took the fort on the hill”

- extract: “Sony took a lesson from the American management books.”

WN: choose, take, select, pick out (pick out, select, or choose from a number of alternatives) “Take any one of these cards”; “Choose a good husband for your daughter”; “She selected a pair of shoes from among the dozen the salesgirl had shown her”;

WN: remove, take, take away, withdraw (remove something concrete, as by lifting, pushing, or taking off, or remove something abstract) “remove a threat”; “remove a wrapper”; “Remove the dirty dishes from the table”; “take the gun from your pocket”; “This machine withdraws heat from the environment”;

WN: take (take into one’s possession) “We are taking an orphan from Romania”; “I’ll take three salmon steaks”;

WN: learn, study, read, take (be a student of a certain subject) “She is reading for the bar exam”.

- view “Principals take cheating seriously.”

WN: take, read (interpret something in a certain way; convey a particular meaning or impression) “I read this address as a satire”; “How should I take this message?”; “You can’t take credit for this!”.

WN: consider, take, deal, look at (take into consideration for exemplifying purposes) “Take the case of China”; “Consider the following case”;

WN: take, submit (accept or undergo, often unwillingly) “We took a pay cut”;

- take the law “She took the law into her own hands”

WN: take (carry out) “take action”; “take steps”; “take vengeance”
4.1 RuleSRL: A Rule-based Semantic Role Labeling System

- take measures “It has taken measures to prevent cheating.”
  \[\text{WN: take (carry out) “take action”; “take steps”; “take vengeance”}\]

- take advantage “.take advantage of the cost-sharing mechanism.”
  \[\text{WN: capitalize, capitalise, take advantage (draw advantages from) “he is capitalizing on her mistake”; “she took advantage of his absence to meet her lover”}\]
  \[\text{WN: trespass, take advantage (make excessive use of) “You are taking advantage of my good will!”; “She is trespassing upon my privacy”}\]

- take some comfort “Investors can take some comfort in the predictable arrival of quarterly dividend checks.”
  \[\text{WN: take (to get into a position of having, e.g., safety, comfort) “take shelter from the storm”}\]

- taking care of business “Frank plans the program, takes care of business, and approaches the work like any other job.”
  \[\text{WN: take care, mind (be in charge of or deal with) “She takes care of all the necessary arrangements”;}\]
  \[\text{WN: attend, take care, look, see (take charge of or deal with) “Could you see about lunch?”; “I must attend to this matter”; “She took care of this business”.}\]

At a quick view, it is obvious that WordNet senses are more refined. Besides the senses presented for the examples above, WordNet has other 29 senses for the verb \text{take}, from more general:

- take, occupy, use up (require (time or space)) “It took three hours to get to work this morning”; “This event occupied a very short time”;

- lead, take, direct, conduct, guide (take somebody somewhere) “We lead him to our chief”; “can you take me to the main entrance?”; “He conducted us to the palace”;

- take, get hold of (get into one’s hands, take physically) “Take a cookie!”; “Can you take this bag, please”.

89
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

to more specific ones:

- film, shoot, take (make a film or photograph of something) “take a scene”; “shoot a movie”;

- contract, take, get (be stricken by an illness, fall victim to an illness) “He got AIDS”; “She came down with pneumonia”; “She took a chill”.

However, one can notice that the expressions take measure, take comfort or take the law into his/her hands are not considered for WordNet senses of the verb take. We can conclude thus that PropBank role sets are rather oriented on expressions, while WordNet senses try to generalize over different situations in order to create the more comprehensive possible synset. Therefore, since the SRL system will have PropBank as the development corpus, it is expected that the semantic roles will be better identified for relative fixed expressions (take advantage, take by surprise, etc).

The Predicate Sense Identification module takes as input the syntactically analyzed file with the predicates identified by the previous module and decides, for each predicational verb or noun, which is the best role set to assign. This module uses as external resources PropBank and NomBank frame files where the role sets are exemplified for each verb, respectively noun; a set of rules for role set identification created by the empirical analysis of the training corpus of annotated English sentences; and a list of frequencies of the assignment of predicate senses in the training corpus. The output of this module is the input file, where each verb or noun previously identified as predicate will have its predicted role set information attached.

If the target predicate has just one role set, that role set is assigned as the predicate sense. If the predicate is a phrasal verb, having the lemma followed by an adverbal particle or preposition in the sentence to be annotated (e.g. take off, start up), the identification of the particular role set is simplified since only the role sets that map the specific predicate-particle pair need to consider for selection.

However, when multiple choices are present, disambiguation methods need to be used in order to select the proper role set, since different role sets (either for the same simple lexical predicate or for a specific predicate-particle pair) may accept different semantic roles. Based on empiric observations over a large set of
4.1 RuleSRL: A Rule-based Semantic Role Labeling System

annotated sentences extracted from PropBank, a set of rules was created in order to select as accurately as possible the proper role set:

- Check if the verb is a simple verb, a phrasal verb or a verbal collocation. Each of these has separate entrance in the PropBank resource;

- Check the PropBank examples within the role sets for adjunct types. Certain verbs appear more frequently with adjuncts than without adjuncts. Also, although the adjuncts are not specific to verbs, but cross-verbal, they can appear preferentially (for instance, for the verb rain, local or temporal adjuncts are more frequent that causal ones). Since the semantic roles are not yet established for the input sentence (the next module will perform semantic role identification), the verb preferences observed in PropBank are compared with the dependency relations that hold between the target predicate in the input sentence and the constituents that act as its dependents (TMP or LOC relations indicate clearly an adjunct). For the ADV dependency relations, the semantic class of the constituent, extracted using WordNet’s hypernyms hierarchy, is also considered, since MALT Parser sometimes prefers to annotate a relation as adverbial, instead of the more refined local or temporal ones. Beside the adjunct/dependency relation types, the lexicalizations of the adjuncts in PropBank are also important, especially the prepositions that introduce them;

- In order to eliminate unnecessary searches through the PropBank and the training corpus, when more role sets are annotated for a specific verb, the type of the roles each role set takes is examined. If all the role sets take the same type/number of roles (except for the adjuncts, which are not specific for a verb\textsuperscript{10}), than the exact sense of the verb is not that important, the one with more annotated examples in the PropBank frame being selected;

- An important source of information for verb sense identification is the frequency of a specific verb role set within the training corpus. This information is used as a backup solution, is case no other method can identify the verb role set.

\textsuperscript{10}The rules are applied in the order of their presentation here, so the information that adjuncts may provide is already considered.
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

4.1.2.3 Semantic Roles Identification

The last module identifies and labels each syntactic dependent of the predicate with its semantic role. The input of this module is a file annotated with syntactic information (part of speech and syntactic dependencies), predicate and predicate role set, and the output is a file where each syntactic dependent of the verb is labeled with its corresponding role. This module uses as external resources PropBank and NomBank frame files; a set of rules for semantic role classification created after analyzing the PropBank annotation guidelines and the training corpus; and a list of frequencies of the assignment of different semantic roles in the training corpus.

The syntactic dependents are considered only on the first level below the predicate. For instance, considering the sentence in Table 4.2 (page 82), the predicate take has as syntactic dependents the preposition from, the noun week, the punctuation sign, and the preposition with (i.e. the words that have as head the verb take). These words are considered the heads of their groups (their group is formed by the words they have as dependents). Thus, by extension, the syntactic dependents of the predicate take will be:

\[
\text{from} \Rightarrow \text{from several vantage points (Prepositional Phrase - PP)} \\
\text{week} \Rightarrow \text{this week (Noun Phrase - NP)} \\
, \Rightarrow , \\
\text{with} \Rightarrow \text{with readings on trade, output, housing and inflation (PP)}
\]

A set of rules was empirically developed in order to assign semantic roles to syntactic dependents, based on the dependency relation between predicate and syntactic constituents, the sense of the predicate and PropBank Annotation guidelines. Examples of these rules are:

- Arg0 for verb predicates is generally the noun group in syntactic relation of subject with the verb for the active voice, or in an object relation with the verb for the passive verb (active or passive verbs are tagged differently by the Stanford Parser);

- Arg1 for verb predicates is usually the direct object;

- Arg1 for noun predicates is the dependent whose group starts with the preposition of, if any;
4.1 RuleSRL: A Rule-based Semantic Role Labeling System

- Relations such as TMP, LOC, MOD, ADV indicate the respective adjuncts: ARGM-TMP (temporal), ARGM-LOC (locative), ARGM-MNR (manner), ARGM-ADV;

- For motion predicates, the preposition from indicates an Arg3 role, while the preposition to indicates the Arg4 role;

- ARGM-REC (reciprocals) are expressed by himself, itself, themselves, together, each other, jointly, both;

- ArgM-EXT indicate the amount of change occurring from an action, and are used mostly for numerical adjuncts like % (raised prices by 15%), quantifiers such as a lot and comparatives such as more (he raised prices more than she did);

- ARGM-NEG is used for elements such as “not”, “n’t”, “never”, “no longer”;

- Only one core argument type is allowed for a specific predicate (only one Arg0 - Arg4). In order to meet this constraint, before the assignment of arguments starts for a sentence, a list of all possible arguments is build. If a constituent is assigned with a numbered argument, the specific argument will be deleted from the list of available arguments. For the non-core roles (adjuncts), this restriction does not hold, since a predicate may contain two or more adjuncts of the same type (it is usually the case of temporal adjuncts);

- In general, if an argument satisfies two core roles, the highest available ranked argument label should be selected, if available, where Arg0 >Arg1 >Arg2 >...>Arg4.

The output of the SRL system for the sentence The economy’s temperature will be taken from several vantage points this week, with readings on trade, output, housing and inflation is graphically presented in figure 4.5.
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

4.1.3 Results

An evaluation metric for semantic roles have been proposed within CoNLL shared task on semantic role labeling (Surdeanu et al., 2008). The semantic frames are evaluated by reducing them to semantic dependencies from the predicate to all its individual arguments. These dependencies are labeled with the labels of the corresponding arguments. Additionally, a semantic dependency from each predicate to a virtual ROOT node is created. The latter dependencies are labeled with the predicate senses. This approach guarantees that the semantic dependency structure forms conceptually a single-rooted, connected (but not necessarily acyclic) graph. More importantly, this strategy implies that if a system assigns the incorrect predicate sense, it still can receive some points for the arguments correctly assigned. For example, for the correct proposition:

\[
\text{verb.01: ARG0, ARG1, ARGM-TMP}
\]

the system that generates the following output for the same argument tokens:

\[
\text{verb.02: ARG0, ARG1, ARGM-LOC}
\]

receives a labeled precision score of 2/4 because two out of four semantic dependencies are incorrect: the ROOT dependency is labeled “02” instead of “01” and the dependency to the “ARGM-TMP” is incorrectly labeled “ARGM-LOC”.

Two types of dependencies relations are evaluated: labeled attachment score and unlabeled attachment score. The labeled dependencies correspond to the total accuracy, where the arguments are correctly attached to the right predicate and correctly labeled with the right semantic role. The unlabeled dependencies correspond to the total accuracy for only correctly identifying the predicate for a semantic role, but misclassifying it as with a wrong semantic role label. For both labeled and unlabeled dependencies, precision (P), recall (R), and F1 scores are computed.
4.1 RuleSRL: A Rule-based Semantic Role Labeling System

Precision and recall scores can also be computed for each sub-task. First, labeled and unlabeled scores for the semantic dependencies to ROOT are of interest. In other words, the unlabeled scores for the ROOT dependencies measure the performance of the predicate identification sub-task. The labeled scores for the ROOT dependencies measure the performance of the predicate identification and predicate sense identification sub-tasks. Additionally, precision and recall scores for non-ROOT semantic dependencies are computed (the performance of the semantic role classification module). These scores are reported for each combination of predicate POS tag and argument label. These statistics indicate how the corresponding system performs for a given argument label from a given corpus (NomBank or PropBank). For example, the label “NN* + A0” indicates the performance for Arg0 arguments from NomBank.

The results of the rule-based SRL system evaluated using the presented metrics are presented in Figure 4.6 for the predicate identification task (bottom) and for the predicate sense evaluation task (top).

<table>
<thead>
<tr>
<th>Precision and recall for LABELED semantic dependencies to ROOT (i.e., predicate identification and classification)</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
</tr>
<tr>
<td>NN</td>
</tr>
<tr>
<td>NNS</td>
</tr>
<tr>
<td>VB</td>
</tr>
<tr>
<td>VBD</td>
</tr>
<tr>
<td>VB2</td>
</tr>
<tr>
<td>VBP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Precision and recall for UNLabeled semantic dependencies to ROOT (i.e., predicate identification)</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
</tr>
<tr>
<td>NN</td>
</tr>
<tr>
<td>NNS</td>
</tr>
<tr>
<td>VB</td>
</tr>
<tr>
<td>VBD</td>
</tr>
<tr>
<td>VB2</td>
</tr>
<tr>
<td>VBP</td>
</tr>
</tbody>
</table>

Figure 4.6: Evaluation of the rule-based SRL System for the Predicate Identification / Predicate Sense Identification module

The results of our rule-based SRL system for the Semantic Role Identification
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

module are presented in Figure 4.7, with results for each pair Predicate part of speech - Semantic role type.

![Figure 4.7: Evaluation of the rule-based SRL System for the Argument Identification task](image)

The rule-based SRL system was evaluated during the ConLL2008 Shared Task competition with a F1 measure of 63.79%. Since the best system participating at the competition scored about 85%, the rule-based method was considered insufficient, since many rules have been missed, or are too particular for a generalization, therefore machine learning techniques have been considered for the development of a better Semantic Role Labeling system, the RuleSRL remaining as a baseline system. However, the rule-based system creation was very useful in identifying the main tasks to be addressed by a semantic role labeling system, as well in better understanding the semantic roles nature. Several rules used in the rule-based system were adapted for the statistic SRL system as post-processing tuning.
4.2 PASRL: a Platform for Adjustable Semantic Role Labeling

Our second approach to semantic role labeling involved the exploration of different machine learning techniques in order to improve the rule-based SRL system, and to check if statistical methods provide different results with regard to semantic role labeling when applied to different languages. In order to easily test different machine learning technologies, with similar implementations and requirements, we used the Weka framework. The developed system turned into a platform for creating Semantic Role Labeling systems for different languages, starting with an annotated corpus. The platform, which will be referred to from now on as PASRL (Platform for Adjustable Semantic Role Labeling), trains different classifiers from the Weka environment using a training corpus of sentences annotated with semantic roles, checks the performances of the different obtained models and selects the best performing model. The 10 fold cross-validation results of all classifiers are also saved since they provide a confusion matrix that can be used to see which classes were correctly predicted by different classifiers. The output of PASRL is a Semantic Role Labeling model that can be applied to a new plain, unlabeled text in order to add semantic role information. The system was developed using training sets for different languages: English, German, Chinese, Czech and Japanese, provided for research purposes by the participation to the CoNLL 2009 shared task.

4.2.1 Weka Classifiers

Machine Learning (ML) methods enable the analysis of large quantities of data in order to automatically predict their annotations. Several standard ML techniques have been incorporated into a software “workbench” called WEKA, for Waikato Environment for Knowledge Analysis. The overall goal of the Weka project is to build a state-of-the-art facility for developing machine learning (ML) techniques and to apply them to real-world data mining problems. The algorithms can either be applied directly to a dataset through a user interface, or be called directly from the code. Weka (presented in the second part of Witten and Frank, 2005) contains tools for data pre-processing, classification, regression, clustering,
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

association rules, and visualization. Written entirely in Java, Weka is an open source software issued under the GNU General Public License.

The classifiers included in the Weka package that were considered for Semantic Role Classification are briefly presented below:

**END** A meta classifier for handling multi-class datasets with 2-class classifiers by building an ensemble of nested dichotomies. (Dong et al., 2005). This classifier uses as base classifier nested Dichotomies (a meta classifier, using as base J48, for handling multi-class datasets with 2-class classifiers by building a random tree structure);

**LogitBoost** Classifier for performing additive logistic regression (Friedman et al., 1998). This class performs classification using a regression scheme as the base learner, and can handle multi-class problems. It can do efficient internal cross-validation to determine the appropriate number of iterations. As base classifier, it uses Decision Stump;

**DecisionTable** Classifier for building and using a simple decision table majority method. (Kohavi, 1995);

**Jrip** This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen in (Cohen, 1995);

**Kstar** K* is an instance-based classifier. The class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function (Cleary and Trigg, 1995). It differs from other instance-based learners in that it uses an entropy-based distance function;

**J48** Class for generating a pruned or unpruned C4.5 decision tree (Quinlan, 1993);

**RandomTree** Class for constructing a tree that considers K randomly chosen attributes at each node. Performs no pruning;

**NaïveBayes** Class for a Naïve Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data (John and Langley, 1995);
4.2 PASRL: a Platform for Adjustable Semantic Role Labeling

ZeroR Class for building and using a 0-R classifier. Predicts the mean (for a numeric class) or the mode (for a nominal class);

LWL - Locally weighted learning. Uses an instance-based algorithm to assign instance weights which are then used by a specified Weighted Instances Handler (Frank et al., 2003). Performs classification using naive Bayes;

IB1 Nearest-neighbor classifier. Uses normalized Euclidean distance to find the training instance closest to the given test instance, and predicts the same class as this training instance (Aha et al., 1991). If multiple instances have the same (smallest) distance to the test instance, the first one found is used;

IBk K-nearest neighbors classifier. Can select appropriate value of K based on cross-validation (Aha et al., 1991);

AttributeSelectedClassifier Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier. Uses as base classifier J48;

SMO Implements John Platt’s (Platt, 1998) sequential minimal optimization algorithm for training a support vector classifier. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. Multi-class problems are solved using pairwise classification;

LibSVM A wrapper class for the libsvm tools. LibSVM reports many useful statistics about LibSVM classifier (e.g., confusion matrix, precision, recall, ROC score, etc.) (Chang and Lin, 2001);

ClassificationViaRegression Class for doing classification using regression methods. The class is binarized and one regression model is built for each class value (Frank et al., 1998). The base classifier used is M5 Model trees and rules (Quinlan, 1992);

ClassificationViaClustering A simple meta-classifier that uses a clusterer for classification. Uses as base SimpleKMeans;
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

RacedIncrementalLogitBoost Classifier for incremental learning of large datasets by way of racing logit-boosted committees, using as base classifier Decision-Stump;

HyperPipes Class implementing a HyperPipe classifier. For each category a HyperPipe is constructed that contains all points of that category (essentially records the attribute bounds observed for each category). Test instances are classified according to the category that “most contains the instance”;

DecisionStump Class for building and using a decision stump. Usually used in conjunction with a boosting algorithm;

RandomForest Class for constructing a forest of random trees (Breiman, 2001);

SimpleCart Class implementing minimal cost-complexity pruning (Breiman et al., 1984).

4.2.2 Training data

The training data used for the development of PASRL was the training and development resource from the ConLL 2009 Shared Task\textsuperscript{11}, consisting of manually annotated treebanks such as the Penn Treebank for English, the Prague Dependency Treebank for Czech and similar treebanks for Catalan, Chinese, German, Japanese and Spanish languages, enriched with semantic relations (such as those captured in the Prop/Nombank and similar resources). The data format is similar to the one presented in tabel 4.1 on page 80.

4.2.3 Learning Steps

Similar to the general architecture of Semantic Role Labeling Systems presented in Section 3.1.3, our system is composed of two main sub-systems: A Predicate Prediction module and an Argument Prediction module (see figure 4.8). The Predicate Prediction module has two possible configurations, corresponding to the Predicate Identification and Predicate Sense Identification layers as marked on the

\textsuperscript{11}ConLL 2009 Shared Task web page: http://ufal.mff.cuni.cz/conll2009-st/
4.2 PASRL: a Platform for Adjustable Semantic Role Labeling

Figure 4.8: Architecture of the PASRL Platform

The first configuration involves a sequential identification of the predicates in a sentence (Predicate Identification modules), followed by the assignment of the predicate sense (Predicate Sense Identification modules on the right side of the figure). The second configuration allows for a joint learning of the predicates in sentence, together with their senses (the Sense Identification modules).

The second sub-system performs argument prediction, based on the dependency relations previously annotated with the MaltParser and the predicate senses (the Argument Identification modules).

For each problem, the modules have three variants: all, NP/VP and each, related to the training set size. Training the whole system for a particular language requires running tens of classifiers, therefore running the classifiers on the whole training size is a very time expensive task. Therefore, the training

---

12See Section 4.1.2.2 for a discussion about the senses of predicates and their marking in the PropBank annotation.
13see Section 4.2.3.4 for more details.
data has been filtered and, besides running the classifiers on the whole data size (all), we offer the possibility to train, for each problem, different classifiers for the noun phrase or verb phrase (NP/VP) or even more refined, for each noun, verb respectively in the training set (each).

For each module in figure 4.8, the classifiers presented in Section 4.2.1 are trained. After running all the classifiers for all the modules, their performance\textsuperscript{14} is compared, and the path in figure 4.8 that obtains the highest performance is considered the best configuration. An example of such a best configuration can be: running the model created by the Predicate Identification – all subtask using the J48 classifier, followed by the model created by the Predicate Sense Identification – NP/VP subtask created with the Decision Table classifier, and then the model of Argument Identification – each subtask created using the Naïve Bayes classifier. The models for this best configuration are saved, and the best path is written to a configuration file. This configuration can then be used at a later time to annotate new texts with the developed SRL system. If all the created models are saved, and not just the best performing ones, the user can define, using the configuration file, the sequence of classifiers it wishes to run for each subtask to annotate new texts using the pre-trained models.

The next sections present each subtask and the results of the best classifiers when evaluated using 10-fold cross-validation on the training data.

4.2.3.1 Predicate Prediction

The first module of the semantic role labeling system is the predicate prediction module. Two approaches are investigated: pipelining the predicate identification with the predicate sense identification, or joint learning the two. This section presents the two approaches, with a discussion on the features used for each step.

\textsuperscript{14}The performance of the classifiers is by default its accuracy, evaluated using 10-fold cross-validation on the training data. The running time of the classifier while testing the created model on the 10 folds is also kept, in case the user requires the best performance to be considered the minimum running time or a combination between the running time and the labeling performance. The selection of the performance type is a parameter of PASRL.
4.2 PASRL: a Platform for Adjustable Semantic Role Labeling

Predicate Identification Task

The input of the Predicate Identification module is the ConLL column-style syntactically and dependency-based parsed format of the training set. The predicate identification program transforms the annotated input into training instances for the ML algorithms in order to identify which nouns or verbs from the input are predicates. For each noun or verb in the sentence, an instance is created with a set of features from a syntax-based vector space, inspired from the features usually used for Semantic Role labeling, presented in Chapter 3, and a binary class label (the candidate word is or not a predicate for the considered sentence). The features used for the Predicate Identification task can be grouped in:

candidate verb-/ candidate noun-dependent features :

- a binary feature establishing if the target word exists in PropBank, respectively NomBank if the candidate is a noun;
- a numeric feature counting the number of dependents;
- a numeric feature counting the number of dependents situated after the candidate predicate;
- a numeric feature counting the number of dependents situated before the candidate predicate;

candidate verb-/ candidate noun-dependent features :

- nominal features representing part of speeches and chunk types of the \( n \) words before and after the candidate word;
- nominal features representing the dependency relation between the candidate word and the words situated up to \( n \) before and after it in the linear order of the sentence, if any;
- binary features representing the is head for relation between the candidate word and the words situated up to \( n \) before and after it in the linear order of the sentence;

context dependent features :

- parent and sibling constituents :

\(^{15}\)The results presented in this Chapter are obtained with 3-grams, but the program is implemented as to allow the user to choose the value of \( n \), the default being three.
• binary features stating if the candidate word has a head that is not in the scope of the n-gram (either after or before);

• binary features stating if the candidate word has other children which are not in the scope of the n-gram (either after or before).

For English, table 4.3 shows the results obtained by different classifiers\(^\text{16}\) for the Predicate Identification task, evaluated using 10-fold cross-validation against a held-out part of the training data.

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>86.323</td>
</tr>
<tr>
<td>ClassificationViaRegression</td>
<td>86.016</td>
</tr>
<tr>
<td>SimpleCart</td>
<td>85.996</td>
</tr>
<tr>
<td>RandomForest</td>
<td>84.366</td>
</tr>
<tr>
<td>AttributeSelectedClass</td>
<td>82.169</td>
</tr>
<tr>
<td>DecisionTable</td>
<td>80.921</td>
</tr>
<tr>
<td>RacedIncrementalLogitB</td>
<td>79.693</td>
</tr>
<tr>
<td>KStar</td>
<td>78.970</td>
</tr>
<tr>
<td>IBk</td>
<td>78.731</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>78.547</td>
</tr>
<tr>
<td>IB1</td>
<td>78.172</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>77.701</td>
</tr>
<tr>
<td>RandomTree</td>
<td>73.902</td>
</tr>
<tr>
<td>LWL</td>
<td>68.227</td>
</tr>
<tr>
<td>DecisionStump</td>
<td>64.488</td>
</tr>
<tr>
<td>LibSVM</td>
<td>63.383</td>
</tr>
<tr>
<td>ClassificationViaClust</td>
<td>57.551</td>
</tr>
</tbody>
</table>

Continued on Next Page...

\(^{16}\)When training the classifiers for the whole data set, there were classifiers that threw Out of Memory error, got stuck, or took longer than the time limit allowed to each classifier to train, so their evaluation is not present in the tables with classifiers results. The time limit was set to about 6 hours per classifier, given that for each module about 20 classifiers can be trained, and there are 12 modules to train and evaluate, but the user can modify it in PASRL parameters.
We observe that, running the classifiers with the default weights of Weka, their results range from 86% correctly identified predicates to 56%. The algorithm that performs best is the J48 classifier and the one that performs worst is the simple ZeroR classifier. Using boosting techniques with J48 as base classifier could improve further the module’s performance. Changing the default weights of the classifiers can modify their performances, but we believe that the hierarchy will not change substantially. However, this remains a direction to address in a further work.

The best performing model (J48 in this case) is saved and will be used when the Predicate Identification module will be called from the configuration file for annotating an unlabeled text.

**Predicate Sense Identification Task**

Since this module considers that the predicational words are already identified, a binary feature *is predicate* is extracted from the training file and added to the feature set created for the Predicate Identification task, and Weka classifiers are again ran to classify each predicate with its PropBank/NomBank role set (e.g. *take.01* for the role set 01 of the predicate *take*).

The results obtained for English are presented in table 4.4 for the Predicate Sense Identification task, evaluated using 10 fold cross-validation against a held-out part of the training data.
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

Table 4.4: ML algorithms evaluated using 10-fold cross-validation for English, for the Predicate Sense Identification task

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogitBoost</td>
<td>61.085</td>
</tr>
<tr>
<td>J48</td>
<td>60.641</td>
</tr>
<tr>
<td>HyperPipes</td>
<td>58.868</td>
</tr>
<tr>
<td>JRip</td>
<td>58.561</td>
</tr>
<tr>
<td>LWL</td>
<td>58.349</td>
</tr>
<tr>
<td>AdaBoostM1</td>
<td>58.322</td>
</tr>
<tr>
<td>DecisionStump</td>
<td>58.322</td>
</tr>
<tr>
<td>AttributeSelectedClass</td>
<td>57.899</td>
</tr>
<tr>
<td>DecisionTable</td>
<td>57.667</td>
</tr>
<tr>
<td>ZeroR</td>
<td>56.548</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>56.064</td>
</tr>
<tr>
<td>IBk</td>
<td>55.621</td>
</tr>
<tr>
<td>KStar</td>
<td>54.577</td>
</tr>
<tr>
<td>IB1</td>
<td>54.345</td>
</tr>
<tr>
<td>ClassificationViaClust</td>
<td>36.937</td>
</tr>
</tbody>
</table>

The results for the Predicate Sense Identification Task presented in table 4.4 are considerably worse than the ones for Predicate Identification task, even if the results report an evaluation performed on a gold annotated input file. Therefore, the actual results are expected to be even worse. However, we notice that J48 is still among the best algorithms, and memory based algorithms generally perform badly on this subtask.

4.2.3.2 Joint Predicate and Predicate Sense Identification

Instead of running the Predicate Identification and the Predicate Sense Identification processes successively, we tested running them simultaneously, using the same features presented in Section 4.2.3.1 and as class the predicate role set similar to the one used for Predicate Sense Identification. The results for English are
4.2 PASRL: a Platform for Adjustable Semantic Role Labeling

presented in table 4.5:

**Table 4.5:** ML algorithms evaluated using 10-fold cross-validation for English, for the Sense Identification task

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>57.974</td>
</tr>
<tr>
<td>AttributeSelectedClass</td>
<td>57.899</td>
</tr>
<tr>
<td>JRip</td>
<td>57.838</td>
</tr>
<tr>
<td>DecisionTable</td>
<td>57.667</td>
</tr>
<tr>
<td>AdaBoostM1</td>
<td>56.548</td>
</tr>
<tr>
<td>DecisionStump</td>
<td>56.548</td>
</tr>
<tr>
<td>Grading</td>
<td>56.548</td>
</tr>
<tr>
<td>HyperPipes</td>
<td>56.542</td>
</tr>
<tr>
<td>IBk</td>
<td>51.200</td>
</tr>
<tr>
<td>KStar</td>
<td>50.743</td>
</tr>
<tr>
<td>RandomTree</td>
<td>49.761</td>
</tr>
<tr>
<td>IB1</td>
<td>48.758</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>46.153</td>
</tr>
<tr>
<td>ClassificationViaClust</td>
<td>35.669</td>
</tr>
</tbody>
</table>

The results show an even worst performance than for Predicate Sense Identification, suggesting that the option of running successively the Predicate Identification, followed by the predicate Sense Identification modules may be better than running the joint task (barely 58%). However, the best classifier is again J48, showing that decision trees are a good algorithm for semantic role labeling (their use has already been accounted for in the literature - see Chapter 3).

### 4.2.3.3 Argument Prediction

After running the first module of the semantic role labeling system (either pipelined or joint), the Argument Identification task is called in order to assign to each syntactic constituent of a verb / noun its corresponding semantic role. The instances
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

in this case are not only the nouns and verbs, but every word in the sentence. The features used are:

- **part of speech** nominal feature with the part of speech the word has;
- **deprel** nominal feature with the dependency relation of the word;
- **chunk** nominal feature with the type of group in which the word is situated (noun group, verb group, prepositional group, punctuation or none);
- **headpos** nominal feature stating what kind of part of speech is the head of the current word;
- **position** nominal feature which establishes if the current word is before, after or exactly the predicate of the sentence\(^\text{17}\);
- **distance to predicate** numeric feature counting the number of words between the current word and the predicate of the sentence;
- **is punctuation** binary attribute that shows if the word is or not a punctuation sign;
- **is son of predicate** binary attribute used to establish if the current word has the predicate of the sentence as head;
- **has head be or modal** binary attribute used to account for the words that have the head a copula or a modal verb;
- **named entity** nominal feature used to establish if the word is a named entity (temporal, person, language, organization, spatial, quantity) or not. The named entities are determined relying on the capitalized writing and WordNet *instance of* function. This feature is only used for English;
- **WordNet hypernym** nominal feature which establishes, if the word is a noun, its WordNet hypernym. A restricted set of hypernyms is considered (spatial, temporal, quantity, object, person, organization, language, entity), and

\(^{17}\text{Each sentence is processed } n \text{ times, where } n \text{ is the number of predicates identified in the sentence, so that at every moment, the words of the sentence are reported to only one predicate.}\)
4.2 PASRL: a Platform for Adjustable Semantic Role Labeling

WordNet’s hierarchy is searched until the first hypernym matches one of the elements in the list\textsuperscript{18}. This feature is only used for English.

**Table 4.6:** ML algorithms evaluated using 10-fold cross-validation for English, for the Argument Identification task

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleCart</td>
<td>95.524</td>
</tr>
<tr>
<td>J48</td>
<td>95.510</td>
</tr>
<tr>
<td>DecisionTable</td>
<td>95.411</td>
</tr>
<tr>
<td>END</td>
<td>95.382</td>
</tr>
<tr>
<td>RandomForest</td>
<td>95.161</td>
</tr>
<tr>
<td>IBk</td>
<td>95.140</td>
</tr>
<tr>
<td>RandomTree</td>
<td>95.140</td>
</tr>
<tr>
<td>JRip</td>
<td>95.081</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>94.849</td>
</tr>
<tr>
<td>AttributeSelectedClass</td>
<td>94.462</td>
</tr>
<tr>
<td>RacedIncrementalLogitB</td>
<td>94.236</td>
</tr>
<tr>
<td>IB1</td>
<td>93.532</td>
</tr>
<tr>
<td>AdaBoostM1</td>
<td>93.132</td>
</tr>
<tr>
<td>DecisionStump</td>
<td>93.132</td>
</tr>
<tr>
<td>HyperPipes</td>
<td>93.132</td>
</tr>
<tr>
<td>ZeroR</td>
<td>93.132</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>88.685</td>
</tr>
</tbody>
</table>

Table 4.6 presents the results obtained after running the Argument identification task on the English data on the correct input, evaluated on the held-out part using cross-validation. The results of almost all the classifiers are very good, but these are the results on gold annotated input. On real data, the performances are expected to be lower.

\textsuperscript{18}Since no word disambiguation module is used, usually the first WordNet sense is the one that imposes its hypernym class, if is not entity (the top class in WordNet). If it is entity, the second sense is also checked, and so on.
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

4.2.3.4 Considerations on data sets

The training data was the data from the CoNLL 2009 Shared task competition. However, the resources were too big to process in one run (over 100,000 sentences), especially since we are running multiple classifiers. Therefore, beside using the whole data set and letting Weka run for hours, we divided the training data in several ways.

NP vs. VP Prediction

A first decision on splitting the training data concerned the different behavior that noun predicates and verb predicates may have, namely:

- noun predicates have fewer arguments than verb predicates, and almost no adjuncts, only core arguments;
- the arguments of the noun predicates are usually much closer (as position in the surface representation of the sentence) to the noun than verb arguments are to the verb;
- there are much fewer role sets for noun predicates than for verb predicates.

Therefore, we allowed the machine learning algorithms to run only on verb, respectively noun candidates at one run. The features have remained the same, the only difference being that instead of running all Weka classifiers on the input data once, we run two times the whole set of Weka classifiers (once for verb and once for noun candidate instances). Even if running twice Weka classifiers may look as no gain in the running time, the smaller instance set (only noun, respectively verb predicates) and the possible parallelization using different threads helps improving the running time. The obtained models can be used for annotation of noun/verb predicate separately. On the Predicate Identification task, the results did not have important differences compared to using the whole dataset, but on the Sense Identification task the difference for the best algorithms were from almost 58% on the whole data to 78% on noun predicates and 74% on verb predicates.
4.2 PASRL: a Platform for Adjustable Semantic Role Labeling

Verb/Noun Specific Prediction

The second and more radical split was training the whole set of classifiers on each verb / noun, if the number of occurrences in the training set was bigger than 10\(^{19}\). The running time substantially diminishes, and the performance of the classifiers increases. However, since a much larger number of classifiers had to be created, the space required on the computer hard drive is increased. Another drawback is that, if the predicate does not exist in the training set or occurs less than 10 times, no model can be created, while when using the NP/VP split for instance, the generalization may help classify a predicate even if it has not been previously seen at all.

Argument Class Prediction

Besides the prediction of arguments on the whole data set, classifiers were trained for each argument type (Arg0, Arg1, \ldots, ARGM-LOC, \ldots). Although the results are improved when using different classifiers for each type, the difference compared to the high performance of Argument Identification on all training data is not tremendous.

4.2.4 Results

The evaluation of the PASRL performance was computed using 10-fold cross-validation on the training set. For each task, PASRL evaluates all the machine learning algorithms used against the gold-annotated corpus, and the best performing algorithm is saved for the configuration file. The evaluation was performed considering the number of correctly classified labels. PASRL was tested using the training data for different languages, available through the ConLL 2009 Shared Task. The 5 best classifiers for the Predicate Identification task on the whole training data, evaluated using 10-fold cross-validation, are presented in the tables below, with their accuracy:

\(^{19}\)The minimum number of instances for each candidate was selected as to conform to the specifications of Weka classifiers.
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

**Table 4.7:** Best 5 ML algorithms evaluated using 10-fold cross-validation for English, considering the whole training set, for the Predicate Identification task

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>86.323</td>
</tr>
<tr>
<td>ClassificationViaRegression</td>
<td>86.016</td>
</tr>
<tr>
<td>SimpleCart</td>
<td>85.996</td>
</tr>
<tr>
<td>RandomForest</td>
<td>84.366</td>
</tr>
<tr>
<td>AttributeSelectedClass</td>
<td>82.169</td>
</tr>
</tbody>
</table>

**Table 4.8:** Best 5 ML algorithms evaluated using 10-fold cross-validation for German, considering the whole training set, for the Predicate Identification task

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdaBoostM1</td>
<td>95.340</td>
</tr>
<tr>
<td>AttributeSelectedClassifier</td>
<td>95.340</td>
</tr>
<tr>
<td>CVParameterSelection</td>
<td>95.340</td>
</tr>
<tr>
<td>DecisionStump</td>
<td>95.340</td>
</tr>
<tr>
<td>END</td>
<td>95.340</td>
</tr>
</tbody>
</table>

**Table 4.9:** Best 5 ML algorithms evaluated using 10-fold cross-validation for Czech, considering the whole training set, for the Predicate Identification task

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleCart</td>
<td>97.789</td>
</tr>
<tr>
<td>ClassificationViaRegression</td>
<td>97.603</td>
</tr>
<tr>
<td>J48</td>
<td>97.582</td>
</tr>
<tr>
<td>END</td>
<td>97.579</td>
</tr>
<tr>
<td>OrdinalClassClassifier</td>
<td>97.579</td>
</tr>
</tbody>
</table>
4.2 PASRL: a Platform for Adjustable Semantic Role Labeling

Table 4.10: Best 5 ML algorithms evaluated using 10-fold cross-validation for Chinese, considering the whole training set, for the Predicate Identification task

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleCart</td>
<td>98.154</td>
</tr>
<tr>
<td>RotationForest</td>
<td>98.116</td>
</tr>
<tr>
<td>ClassificationViaRegression</td>
<td>98.059</td>
</tr>
<tr>
<td>FilteredClassifier</td>
<td>98.029</td>
</tr>
<tr>
<td>J48</td>
<td>98.010</td>
</tr>
</tbody>
</table>

Table 4.11: Best 5 ML algorithms evaluated using 10-fold cross-validation for Japanese, considering the whole training set, for the Predicate Identification task

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimpleCart</td>
<td>79.883</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>79.367</td>
</tr>
<tr>
<td>ClassificationViaRegression</td>
<td>79.333</td>
</tr>
<tr>
<td>LADTree</td>
<td>79.333</td>
</tr>
<tr>
<td>J48</td>
<td>79.023</td>
</tr>
</tbody>
</table>

We can observe that SimpleCart, J48 and ClassificationViaRegression are among the 5 best algorithms for different languages, which suggests that semantic role labeling may not be very language dependent, once the input text is syntactically annotated, which seems natural since semantic roles are representations at the deep, conceptual structure of the language. This observation supported our intuition that semantic roles are likely to be valid cross-linguistically, which stayed at the basis of the creation of the Romanian Resource of Semantic Roles (see Chapter 5).

Another important observation is that the best performing algorithms score differently for different languages. For instance, the best performances for Czech, German and Chinese range around 97-98%, while for English the best perfor-
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM

Performance is 86% and for Japanese only almost 80%. Since the size of the training corpus is similar, one possible explanation could be the different level of inflection, the free vs. fixed word order, or the position of the verb with respect to the other elements of the sentence (SVO vs. SOV language type), or a combination thereof.

4.2.5 Using PASRL pre-trained models

After PASRL is trained, it can be used to annotate raw text with the models saved. A configuration file is used to tell the program which models will be run. The best models created during the development phase of PASRL are offered as pre-trained models, if the user only wants to use PASRL as a semantic role labeler, and not a semantic role labeling model creator. The plain text is pre-processed using the same tools as in the case of the RuleSRL system presented in section 4.1: Stanford Parser for Part-of-speech identification and MaltParser for dependency relations. For languages other than English, the MaltParser can be trained on the training corpus provided by the CoNLL 2009 Shared Task organizers. However, Stanford Parser has build-in language models for English, German and Chinese, for the other languages, a language-specific POS-tagger is needed.

When evaluating the pre-trained models for English on new data, using the whole processing chain (including part of speech and dependency annotation), the results are promising, with 68% for noun predicate and 81% for verb predicate F1.

4.2.6 Using PASRL for Romanian

In order to use PASRL platform to create semantic role labelers for the Romanian language, a corpus of annotated examples in Romanian is needed to be used as training dataset. Chapter 5 presents the creation of such a resource for Romanian, containing a set of 1500 sentences. PASRL is run and the best configuration for the Romanian language is found to be: Predicate Identification all with J48, Predicate Sense Identification NP/VP with SimpleCart, Argument Identification all with J48 and the evaluation of the overall performance of the system using cross-validation on the training set shows an accuracy of 67%. The results are
4.3 Conclusions

promising, and we believe that they can be considerably improved by using a larger dataset as training corpus. The envisaged solution supposes using the collection of juridical multilingual parallel documents of the European Union (the JRC Acquis 20) in order to create (using the method described in Chapter 5) a bigger Romanian role resource.

To annotate a new Romanian text with semantic roles using the pre-trained models of PASRL, we add the syntactic information using RACAI web service 21 and a dependency parser developed training the MALT parser on a corpus of Romanian sentences annotated with dependency information (Moruz et al., 2006). The performance of annotating new Romanian sentences with the Romanian pre-trained PASRL modules has not yet been accurately computed, but the first estimations show promising results.

4.3 Conclusions

This chapter has presented the development of a rule-based Semantic Role Labeling system aiming to annotate raw text with semantic roles. The input text is initially pre-processed by adding part of speech information using the Stanford Parser and syntactic dependencies using the MaltParser. The modest results of the rule-based system led us to consider applying machine learning techniques. Different machine learning algorithms have been used for Semantic Role Labeling (SRL) over the time (see Chapter 3). In order to investigate which one is best suited for the SRL task, we have developed a “platform” for creating supervised Semantic Role Labeling systems starting with an annotated corpus and using the algorithms implemented in Weka framework. The developed platform tests several classifiers on different sub-problems of the SRL task (Predicate Identification, Predicate Sense Identification, Sense Identification, Argument Identification), chooses the ones with the greatest performance and returns a Semantic Role Labeling System (a sequence of trained models to run on new data). Evaluations of the performances of the best algorithms for several languages are discussed.

---

21] RACAI web services: http://www.racai.ro/webservices
4. DEVELOPMENT OF A SEMANTIC ROLE LABELING SYSTEM
Romanian Resource of Semantic Roles

The realization of human-computer interaction in natural language represents a major challenge in the context of aligning Romanian to existing technologies. The proposed thesis aims to introduce the semantic frames to Romanian NLP systems.

5.1 Intuition

Annotated language resources have become a must in natural language processing, especially for supervised learning (training and evaluation), unsupervised learning (evaluation), hand-crafted systems (evaluation), etc. Quality control is an important issue, since annotations, in order to be used as gold standard for evaluation, need to be very accurate. Inter-annotator agreement metrics have been developed (an overview is presented by Artstein and Poesio, 2005), but the major problems remain the temporal, financial and human resources needed in order to ensure a (near) perfect corpus. What if we have short deadlines and limited human and financial possibilities? Could we use existing language resources, built with considerable efforts for a specific language, and import them for a new language? We believe this is feasible, and this chapter exploits the application of this method on building a semantic role resource for Romanian through the import of the English FrameNet annotation.
Pado and Erk (2005) have performed a manual pilot study investigating the use of frame semantics for parallel semantic analysis of aligned text. Looking at translation pairs differing in their parts of speech, they found that predicate-argument structure abstracts somewhat from morphologic and syntactic language idiosyncrasies, but there are still considerable variations in the distribution of semantic material over predicates. They propose an algorithm which derives frame paraphrases (which are matched to semantic substructures rather than word sequences) from a bilingual corpus. The algorithm provides a means of identifying and systematically studying cross-lingual mismatches in the distribution of semantic material. The resulting data allows an analysis of semantic differences e.g. in the expression of different degrees of causality, although the algorithm was applied only manually to a small dataset. First, the application of the algorithm to the translation pair “increase-höher” offers a corpus-based view on expressions on a continuum between causation and non-causation and on the conditions of their use. Secondly, the study has provided a first impression of the degree of cross-lingual uniformity of frame-semantic structure.

Creating from scratch a semantic role resource implies several steps before the annotation process itself: (1) finding a corpus, (2) establishing an annotation schema and defining annotation guidelines, (3) choosing/creating an annotation software, (4) training annotators. The basic daily routine of semantic role resource annotators is defined in (Baker and Sato, 2003) for FrameNet, and involves:

- define a frame and its roles;
- make a list of words that evoke the frame (its lexical units LUs);
- extract example sentences containing these LUs from a corpus;
- semi-automatically annotate the parts of the sentences which are the realizations of these roles, including marking the phrase type (PT) and grammatical function (GF);
- automatically create a report which constitutes a lexical entry for this LU, detailing all the possible ways in which the roles can be syntactically realized.
The FN software used for frame definition and annotation (Baker and Sato, 2003; Fillmore et al., 2002), combines the frame editing tools and the annotation tools into a single GUI, in a MySQL client-server model. With different corpora, annotation schema or software, the procedure is similar for all semantic roles resources manually built. Though the resource becomes a gold standard by its correctness, the time needed to develop such resources extends to several years and the process requires hundreds of annotators.

Creating a semantic role resource for a new language (Erk et al., 2003; Subirats-Ruggeberg and Petruck, 2003) is usually performed similar to the way the English semantic roles resource was created. The main difference is that the annotation schema is already established. However, it still takes considerable time to go through the annotation process on the new corpus. To estimate the time needed to create a semantic role resource for Romanian, we considered that we have a generic Romanian corpus, that we will use the English FrameNet annotation schema and software, that we have two very well trained annotators, with good semantic frames knowledge, and that we only need to worry about the annotation process itself. Our tests revealed that a person can (optimistically) annotate an average of 45 medium sized sentences per hour. For a target of 100,000 sentences, we computed around 2200 hours, i.e. 13 months, considering 8 hours a day, 5 days a week working time. The main problem with this approach was the lack of a definite list of possible semantic roles. Therefore, different annotators can give different names to the same role (Agent or Seller or Vendor, for instance, can be the doer of the selling process, the person that alienates the goods), confusing the corpus quality metrics and the inter-annotator agreement.

The starting point for the German, Japanese and Spanish FN creation was the manual annotation at FE level of existing corpora for each language. To save time, the approach we adopted was the direct import of the English annotation to Romanian by translating the sentences from the English annotated resource. The intuition behind our import program, detailed in (Trandabăt, 2007), is that most of the frames defined in the English FN are likely to be valid cross-linguistically, because semantic frames express conceptual structures, language independent, at the deep structure level. Several examples to support our intuition are presented below:
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

(1)  
EN: ...until [Craig McDermott]_{Entity} becomesTARGET 
RO: ...până când [Craig McDermott]_{Entity} va deveniTARGET 
[available]_{Final,tate} in 1994_{Time}. 
[disponibil]_{Final,tate} în 1994_{Time}. 

(2)  
EN: [By the end of the eighteenth century]_{Time}, [Lane]_{Entity} 
RO: [Pe la sfârșitul secolului optsprezece]_{Time}, [Lane]_{Entity} 
[had become]TARGET [wealthy]_{Final,tate} 
[devenise]TARGET [bogat]_{Final,tate} 
[by profits from them and from military activities]_{Cause}. 
[din profituri și din activități militare]_{Cause}. 

With the development of word alignments methods for parallel corpora, the idea of using one languages annotation to induce an annotated resource for another language has come into view. Yarowsky et al. (2001) have described a system and a set of algorithms for automatically inducing stand-alone monolingual part-of-speech taggers, named-entity taggers and morphological analyzers from English to French, Chinese, Czech and Spanish. The assumption that for two sentences in parallel translation, the syntactic relationships in one language directly map to the syntactic relationships in another language (named the direct correspondence assumption) has also been studied by Hwa et al. (2002). They provided a Direct Projection Algorithm for transferring English syntactic relations to Chinese. For the Romanian language, Barbu Mititelu and Ion (2005) have analyzed the import of verbal dependency relations on a word-aligned parallel English-Romanian corpus. After the analysis of the syntactic relations transfer, the semantic relations received the researchers attention, especially after the development of large English resources. The transfer of semantic information from one language to another has started to be considered for WordNet sense mapping (Bentivogli and Pianta, 2002; Lupu et al., 2005) and anaphora resolution (Postolache et al., 2006). At the “Romance FrameNet” Workshop within Eurolan 2005 Summer School\(^1\), two papers have begun to investigate the transfer of semantic relations from English FrameNet to Spanish (Johansson and Nugues, 2005) and to Romanian (Trandabăț et al., 2005) using word-level alignment of parallel corpora.

The semantic concept level is mostly cross-linguistically constant, the surface

\(^1\) Romance FrameNet Workshop web page: http://www.icsi.berkeley.edu/ vincenzo/rfn/index.html
realization of these concepts and their relations (the syntactic constraints) being changed from one language to another. However, we must be careful that the English-centric approach could endanger the realization of a true language semantic FrameNet, tempting the exact copying of the English semantic frames, although inter-lingual differences may be actually manifested. After all, the transfer starts from lexical predicates, and lexicals are language dependent.

The steps required in importing semantic roles from English to Romanian are: (1) select an English annotated sentence, (2) translate it to Romanian, (3) align the English and the Romanian sentence at word level, (4) transfer the annotation from English to Romanian, (5) validate and correct the import. The import program is described in the next section.

For the import method, the main time consuming task is the translation. Although the translation rate may be similar to the annotation rate (around 40-45 sentences per hour), the real gain is that the corpus can be split to several translators (cheaper and easier to find than semantic annotators), thus finishing the translation in about three months. After the automatic alignment and import, a single annotator is needed to perform the validation of the created corpus, focusing on cases where the alignment was not 1:1. Considering that we don’t have two annotators to work for one year just on semantic annotation, and the belief that, once the import program is finished, every other language could benefit from it and transfer annotations for its own language, we created a Romanian FrameNet based on the English annotation.

5.2 System Architecture

The automatically importing program (Trandabăt, 2007) is based on the correlation of the semantic roles expressed in English with the translation equivalents in Romanian of the words that realize a specific role. Figure 5.1 presents the proposed architecture of the semantic roles import system.

In order to test the annotation import method, the first step was the manual translation of as small subset of 110 randomly selected sentences from the English FN. In order to align the Romanian version with the English one, a larger corpus was needed, so the translation continued with the Event frame, summing up to 1094 sentences for the development phase of the project. After the selection of the
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

Figure 5.1: Architecture of the system used to transfer English semantic role annotations to Romanian sentences and their translation, the Romanian sentences were aligned with the English ones using the aligner developed by the Institute of Research in Artificial Intelligence (presented in Tufiş et al., 2005). The next step was the automatic import of the English annotation, followed by a manual validation, a detection of the mismatching cases and an optimization process which, based on inference rules, corrects the automatic annotation.

The importing program was written in Perl, having also an user interface, created to help annotators validate the import process. The interface allows the user to perform the following tasks, related to the import of semantic roles:

- load sentences from an annotated English frame;
- translate them;
- align the English and Romanian versions;
5.2 System Architecture

- import the roles;
- visualize, correct and save the annotation files.

The next sections present each module of the import program, with examples for the English sentence:

The incident occurred after a dispute between the man and staff at a branch of the Bank of Ireland in Cahir.

5.2.1 The Translation Module

The first step in importing the alignment of the English semantic role annotation to Romanian consists in the translation of the English annotated sentence. The sentence is extracted from its annotation frame, which is an XML file with a structure similar to the one presented in Figure 5.2 for the sentence The incident occurred after a dispute between the man and staff at a branch of the Bank of Ireland in Cahir.

![Figure 5.2: FrameNet Semantic Role Annotation Format](image)

In figure 5.2, the *layers* tag contains the annotation of an English sentence: the semantic roles (FE frame elements) Event, Time and Place, and the target predicational verb occurred. The *label* tag contains a name a, an ID, a start and
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

an end position (character) for the sentence substring that matches the specific role. For instance, the role Event start at position 0 and ends at position 11, corresponding to the noun group The incident.

For the development phase of the system, about 1000 sentences were extracted from the English XML annotated frames and manually translated. Though importing annotations for 1094 sentence was a very good start, in order to create a Romanian semantic role resource suitable to machine learning techniques, more annotation imports are needed. To facilitate the use of the import program for large scale annotation transfer, an interface was created. The interface loads the English sentences from the annotated XML files, asking the user to translate each sentence. Beside the manual translation method, due to the fast development of free Internet translation services, we added to the interface an option to automatically translate the sentence using Google translation service. The returned translation may be corrected and saved for further use. Figure 5.3 shows the translation obtained with Google translation service for the sentence considered for exemplification throughout this section. In the interface in figure 5.3, the right column represents the list of English annotated sentences from FrameNet, ordered by their ID, the left top text box lists the English sentence, while the bottom text box contains the translation of the English sentence into Romanian using Google Translation service.

After correcting the sentence (or translating it directly manually), we obtain the following Romanian translation:

Incidentul a apărut după o dispută între individ și personal la o filială a Băncii Irlandeze din Cahir.

From a formal point of view, the English sentence can be viewed as a set of ordered tokens, with punctuation being also treated as a token:

\[ S_{en} = (w_{e1}, w_{e2}, \ldots, w_{en}) \]  \hspace{2cm} (5.1)

and the Romanian translation can be considered:

\[ S_{ro} = (w_{r1}, w_{r2}, \ldots, w_{rm}) \]  \hspace{2cm} (5.2)
5.2 System Architecture

5.2.2 Alignment Module

Most word-level aligners are model-estimating aligners, relying on the IBM models (1 to 5) described in (Brown et al., 1993). One of the first reliable and publicly available implementation is GIZA (Och and Ney, 2000), part of the SMT toolkit EGYPT (Al-Onaizan et al., 1999), developed by the Statistical Machine Translation team at the Center for Language and Speech Processing at Johns-Hopkins University. The program was later improved by Och and Ney and released as GIZA++. GIZA++ implements IBM models with a dependency of word classes as described in (Brown et al., 1993), HMM alignment model with Baum-Welch training, Forward-Backward algorithm, empty word, dependency on word classes, transfer to fertility models, as well as perplexity calculation.

For the development phase of the import program, the English sentences have
been aligned at word level using the RACAI’s\textsuperscript{2} aligner COWAL described in (Tufiş et al., 2005). The Combined Word Aligner, COWAL, is a wrapper of two aligners: a translation equivalence extraction program TREQ-AL (Tufiş, 2002) and MEBA, an iterative algorithm which uses the translation probabilities, distortions and POS-affinities generated by GIZA++. It is complemented by a graphical user interface that allows for the visualization of the alignments (intermediary and the final ones) as well as for their editing.

Although COWAL has an alignment precision of more than 87%, since the RACAI Aligner does not have yet a web service interface, and due to the fact that the import program is supposed to work also unassisted, the alignment have been implemented to work only with GIZA++.

As an example, table 5.1 presents the alignment results using the RACAI Aligner for the sentence used as example through this section, presented in two column format\textsuperscript{3}.

The colored lines in the table mark the alignment errors:

1. the Romanian article “o” in “la o filială” was not aligned with the English article “a” in “at a branch”

2. the Romanian possessive article “a” in “filială a Băncii” not recognized as a translation of the English “of” in “branch of the Bank”, all of these words being aligned to 0.

GIZA++ alignments are of modest accuracy for Romanian comparing to the aligner developed at RACAI, tuned for the Romanian language, and manual validation is required. The system can work also without any correction, but the import results are limited to the translation and alignment performance. In order to correct the automatic alignment offered by GIZA++, an alignment interface was created, where for each English word, the Romanian counterparts considered to be its alignment are highlighted. The interface offers the user the possibility to change and save the corrected alignment.

\textsuperscript{2}Research Institute for Artificial Intelligence of the Romanian Academy

\textsuperscript{3}The alignment program returns a set of indexes indicating the positions of the words in the English, respectively Romanian, sentence, starting at 1 for the first word. For expressiveness reasons, the indexes were replaced with the words situated at the corresponding positions, except for the alignment to 0, which means that the word has no translation equivalent in the other language.
### 5.2 System Architecture

<table>
<thead>
<tr>
<th>Romanian</th>
<th>English</th>
<th>Romanian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidentul</td>
<td>The</td>
<td>o</td>
<td>0</td>
</tr>
<tr>
<td>Incidentul</td>
<td>incident</td>
<td>0</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>occurred</td>
<td>filială</td>
<td>branch</td>
</tr>
<tr>
<td>apărut</td>
<td>occurred</td>
<td>a</td>
<td>0</td>
</tr>
<tr>
<td>după</td>
<td>after</td>
<td>0</td>
<td>of</td>
</tr>
<tr>
<td>o</td>
<td>a</td>
<td>Băncii</td>
<td>the</td>
</tr>
<tr>
<td>dispută</td>
<td>dispute</td>
<td>Băncii</td>
<td>Bank</td>
</tr>
<tr>
<td>între</td>
<td>between</td>
<td>0</td>
<td>of</td>
</tr>
<tr>
<td>individ</td>
<td>the</td>
<td>Irlandeze</td>
<td>Ireland</td>
</tr>
<tr>
<td>individ</td>
<td>man</td>
<td>din</td>
<td>in</td>
</tr>
<tr>
<td>şi</td>
<td>and</td>
<td>Cahir</td>
<td>Cahir</td>
</tr>
<tr>
<td>personal</td>
<td>staff</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>la</td>
<td>at</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

In order to consider the alignment of the English words a surjective function, i.e. there exists at least one English word for each Romanian word, we added the null element to both the English and Romanian sentences, thus the alignment function looks like:

\[ \text{Align}(w_e) : S_{en} \cup \{\emptyset\} \mapsto 2^{S_{ro}} \cup \{\emptyset\} \]

As an example, for the English word “occurred” we have:

\[ \text{Align}(\text{occurred}) = (a \text{ apărut}) \]

while the English word “of” aligns to null.

For the import module, an important information will be the number of aligned correspondences a specific word has. Thus, we define \#Alignments\(w_e\) as the cardinality of the set of alignment for a specific word, \(\text{Align}(w_e)\) or \(W^I_R\):

\[ \#\text{Alignments}(w_e) = |\text{Align}(w_e)| = |W^I_R| \]

with \(W^I_R \subset S_{ro} \cup \{\emptyset\}\).

5.2.3 Import Module

The Semantic Roles Import program is based on the assumption, discussed in section 5.1, that, if a word is part of a specific role in English, it will be part of the same role type in Romanian. The import program uses as input files the annotations of the English lexical units and the aligner’s output files. The import algorithm is very simple, but effective, focusing on:

- reading the English XML files and the alignment files;
- linking each English word with its corresponding semantic role (FE);
- mapping the English words with the aligned Romanian translation, thus transferring the annotation of a specific role from English to Romanian. The mapping was performed by considering the import as a sequential labeling problem, with a B-I-O encoding.
5.2 System Architecture

In the B-I-O representation, a word is characterized as being at the **Beginning**, **Inside** or **Outside** of a sequence to be analyzed, in our case, of a semantic frame. As an example, the English sentence:

> [The incident]_{Event} occurred_{TARGET} [after a dispute between the man and staff]_{Time/Cause} [at a branch of the Bank of Ireland in Cahir]_{Place},

can be seen as:

> The_{B_Event} incident_{I_Event} occurred_{B_TARGET} after_{B_Time/Cause} a_{I_Time/Cause} dispute_{I_Time/Cause} between_{I_Time/Cause} the_{I_Time/Cause} man_{I_Time/Cause} and_{I_Time/Cause} staff_{I_Time/Cause} at_{B_Place} a_{I_Place} branch_{I_Place} of_{I_Place} the_{I_Place} Bank_{I_Place} of_{I_Place} Ireland_{I_Place} in_{B_Place} Cahir_{I_Place} \_O_NO-Frame

The name of the frame is kept after the B-I-O symbols (*e.g.* I_Place, B_Event, etc.), and for the words outside a frame, the NO-Frame name is used (_O_NO-Frame).

Returning to the formal representation, we assign to the English words their annotated frames through \( Frame(w_i) \):

\[
\forall w_i \in S_{en}, \ Frame(w_i) : \{S_{en} \cup \emptyset\} \mapsto \{B_iF_i, I_iF_i, O_iF_i\}
\]

with \( F_i \in \{\cup \ FrameNetRoles, \ NO-Frame\} \)

\[\text{Frame(\emptyset)} = \{\emptyset\}\]

The assignment of \( \text{Frame(\emptyset)} \) is used for the null element introduced in the English and Romanian sentences to cope with the case when words have no alignment in the other language.

For the import method, given the English sentence \( S_{en} \), the Romanian sentence \( S_{ro} \), and the \( \text{Align} \) mapping function, for \( \forall w_i \in S_{en} \), \( \text{Align}(w_i) = W_R^i \subset S_{ro} \cup \{\emptyset\} \), we have the following cases:

**one-to-zero** alignment, when there \( \exists \) an unique \( w_i \in S_{en} \), such that \( \text{Align}(w_i) = \emptyset \), meaning that an English word has no Romanian correspondent;

**one-to-one** alignment, when there \( \exists \) an unique \( w_i \in S_{en} \) and an unique \( w_j \in S_{ro} \), such that \( \text{Align}(w_i) = w_j \), with \( |W_R^i| = 1 \), meaning that an English word has one and only one Romanian correspondent;
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

one-to-many alignment, when there ∃ an unique \( w_{e_i} \in S_{en} \) and a subset of Romanian words \( \{w_{r_j}, \ldots w_{r_l}\} \in 2^{S_{ro}} \), such that \( \text{Align}(w_{e_i}) = \{w_{r_j}, \ldots w_{r_l}\} \), with \( |W'_R| > 1 \), meaning that an English word is translated with more than one Romanian words;

many-to-one alignment, when there ∃ a subset \( \{w_{e_i}, \ldots w_{e_k}\} \in 2^{S_{en}} \) and an unique \( w_{r_j} \in S_{ro} \), such that \( \text{Align}(w_{e_i}, \ldots w_{e_k}) = w_{r_j} \), with \( |W'_R| = 1 \), meaning that several English words are translated with the same Romanian word;

many-to-zero alignment, when there ∃ a subset \( \{w_{e_i}, \ldots w_{e_k}\} \in 2^{S_{en}} \), such that \( \text{Align}(w_{e_i}, \ldots w_{e_k}) = \emptyset \), with \( |W'_R| = 0 \), meaning that several English words are translated with the same Romanian word. This is reduced to one-to-zero alignment, applied for each of the English words in the alignment set which maps to null;

many-to-many alignment, when there ∃ a subset \( \{w_{e_i}, \ldots w_{e_k}\} \in 2^{S_{en}} \) and a subset \( \{w_{r_j}, \ldots w_{r_l}\} \in 2^{S_{ro}} \), such that \( \text{Align}(w_{e_i}, \ldots w_{e_k}) = \{w_{r_j}, \ldots w_{r_l}\} \) and \( |W'_R| > 1 \). This case is mostly theoretical, since in the alignments of the 1094 sentences used to create the initial Romanian Semantic Role Resource, no such alignment was found.

zero-to-one alignment, when for \( w_{e_i} = \emptyset \) there ∃ an unique \( w_{r_j} \in S_{ro} \), such that \( \text{Align}(w_{e_i}) = w_{r_j} \), meaning that there are Romanian words that correspond to no English words (usually it is the case for functional words, introduced to keep the syntactic function of the sentence;

zero-to-many alignment, when for \( w_{e_i} = \emptyset \in S_{en} \), there ∃ a subset \( \{w_{r_j}, \ldots w_{r_l}\} \in 2^{S_{ro}} \), such that \( \text{Align}(w_{e_i}) = \{w_{r_j}, \ldots w_{r_l}\} \) and \( |W'_R| > 1 \). This situation is reduced to zero-to-one alignment, applied for each Romanian word;

zero-to-zero alignment, when \( w_{e_i} = \emptyset \in S_{en} \), \( \text{Align}(w_{e_i}) = \emptyset \). This is not a real alignment case, since no English word is mapped to no Romanian word. This case is presented only for the symmetry of the presentation of alignment cases.

The import strategy used for these cases is detailed in the next sections.
5.2 System Architecture

5.2.3.1 One-to-zero import

In the case of one-to-zero alignment, there $\exists$ an unique $w_{ei} \in S_{en}$, such that $\text{Align}(w_{ei}) = \emptyset$, meaning that the considered English word has no Romanian correspondent. In this case, there is no Romanian word to import the English frame to, therefore no action is needed.

5.2.3.2 One-to-one import

In the case of one-to-one alignment, there $\exists$ an unique $w_{ei} \in S_{en}$ and an unique $w_{rj} \in S_{ro}$, such that $\text{Align}(w_{ei}) = w_{rj}$, with $|W_R^i| = 1$, meaning that the considered English word has only one Romanian word as translation. In this case, the semantic frame annotated for the English word will be transferred to the Romanian word:

$$\text{Frame}(w_{rj}) = \text{Frame}(w_{ei})$$

5.2.3.3 One-to-many import

One-to-many alignments are the cases where there $\exists$ an unique $w_{ei} \in S_{en}$ and a subsetset of Romanian words $\langle w_{rj} \ldots w_{rl} \rangle \in 2^{S_{ro}}$, such that $\text{Align}(w_{ei}) = \langle w_{rj} \ldots w_{rl} \rangle$, with $|W_R^i| > 1$, one English word being translated with two or more Romanian words. As an example, we have the English verb “occurred” aligned with the Romanian verbal group “a apărut” in table 5.1. In this situation, we have to split one English label for two or more Romanian words. Based on empiric observations and the semantic roles nature, we decided to apply the following rule:

$$\text{Frame}(w_{rj}) = \begin{cases} \text{Frame}(w_{ei}), & \text{if } \text{Frame}(w_{ei}) = \{B.F_i\} \text{ and we are at the first Romanian word in the alignment set } \text{Align}(w_{ei}) \\ \text{Frame}(w_{ei}), & \text{if } \text{Frame}(w_{ei}) \in \{I.F_i,O.F_i\} \\ I.F_i, & \text{otherwise} \end{cases}$$

For the example of the verb “occurred $B._{TARGET}$”, having the alignment set $\text{Align}(\text{occurred}) = \{a, \text{apărut}\}$, the English frame $B._{TARGET}$ is imported to the first Romanian word in the alignment set, “a”, and the other words in the alignment set receive $I._{TARGET}$, issuing the imported annotation $a._B._{TARGET}$ $\text{apărut}_I._{TARGET}$. 

131
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

5.2.3.4 Many-to-one import

The opposite situation occurs when two or more English words align with the same Romanian word. In this case, there ∃ a subset \( \{w_{e_1}, \ldots, w_{e_k}\} \in 2^{S_{en}} \) and an unique \( w_{r_j} \in S_{ro} \), such that \( \text{Align}(w_{e_1}, \ldots, w_{e_k}) = w_{r_j}, \) with \( |W_R| = 1 \). This is especially true for the definite article, which in English is placed before the determined noun, while in Romanian it has a pro-clitic position, being merged to the noun. In this case, as opposed to the previous situation, we have to merge two annotations and assign only one frame to only one Romanian word. If the English words belong to the same frame (e.g. all the English words are marked with \( B_{.F_i} \), or all the words are \( I_{.F_i} \)), the merge result is simply the English frame. If the English frame name are the same, but with different B-I-O symbols, a precedence rule \( B > I > O \) is applied, and the frame name with the higher position in the precedence rule is selected. For instance, the English noun phrase “The incident” maps to only one Romanian word, “Incidentul”. In this case, the English annotation to be imported is \( \text{The}\_B\_Event \) and \( \text{incident}\_I\_Event \). Due to the precedence rule, the imported frame is going to be:

\[
\text{The}\_B\_Event \text{ incident}\_I\_Event \implies \text{Incidentul}\_B\_Event
\]

If the English frame names are not the same, but the B-I-O symbols are, for instance we have \( B_{.F_i} \) and \( B_{.F_k} \), the following rule applies to decide which frame name to choose from the pair \( (F_i, F_k) \):

\[
\text{Frame}(w_{r_j}) = \begin{cases} 
F_i, & \text{if } F_i = \text{TARGET, or } F_k = \text{NO-Frame} \\
F_k, & \text{otherwise}
\end{cases}
\]

The same rule applies, correlated with the precedence rule, if we have don’t have frame name, nor B-I-O symbols similarities between the frames to be aligned to \( w_{r_j} \).

5.2.3.5 Zero-to-one import

In this case for \( w_{e_i} = \emptyset \) there ∃ an unique \( w_{r_j} \in S_{ro} \), such that \( \text{Align}(w_{e_i}) = w_{r_j} \). This case is the most problematic one, since it requires introducing a frame annotation to the Romanian word, without having a frame annotation in English, since the Romanian word has no English correspondent. However, we can say
with big confidence that a Romanian word needs to have an \( I.F_i \) annotation when, by looking at the English context, we notice that the word interrupts a frame (it breaks a consecutive sequence of \( B.F_i \) and \( I.F_i \) or a sequence of \( I.F_i \)'s). Otherwise, the Romanian word will have the frame \( O.NO-Frame \), since no automatic decision can be made about the name of its frame:

\[
\text{if } w_{e_i} = \emptyset \text{ there } \exists \text{ an unique } w_{r_j} \in S_{ro}, \text{ such that } \text{Align}(w_{e_i}) = w_{r_j}
\]

\[
\text{Frame}(w_{r_j}) = \begin{cases} 
I.F_i & \text{if } \text{Frame}(w_{r_{j-1}}) = B.F_i \text{ or } I.F_i \text{ and } \\
\text{Frame}(w_{r_{j+1}}) = I.F_i & \\
O.NO-Frame & \text{otherwise}
\end{cases}
\]

For the sentence used as example for the import system throughout this Chapter, we get the final imported Romanian annotation:

\[
\text{Incidentul}_B \text{Event}_B \text{ a}_B \text{TARGET}_B \text{ apărut}_I \text{TARGET}_I \text{ după}_B \text{Time/Cause}_I \text{ o}_I \text{ Time/Cause}_I \text{ dispută}_I \text{Time/Cause}_I \text{ între}_I \text{Time/Cause}_I \text{ individ}_I \text{ _Time/Cause}_I \text{ şi}_I \text{Time/Cause}_I \text{ personal}_I \text{Time/Cause}_I \text{ la}_B \text{Place}_I \text{ a}_I \text{Place}_I \text{ filială}_I \text{Place}_I \text{ a}_I \text{Place}_I \text{ Băncii}_I \text{Place}_I \text{ Irlandeze}_I \text{Place}_I \text{ din}_I \text{Place}_I \text{ Cahir}_I \text{Place}_I \_O.NO-Frame
\]

\section*{5.2.4 Validation Module}

An import validation interface is also provided, in order to allow the annotator to correct if the import rules are not appropriate, or if syntactic constraints force the semantic frames to be different in the two languages. In the case of the sentence that served as example in this section, the import was correctly performed by the system, although the alignment contains some mismatches, discussed above. The final annotation for the translated sentence will be:

\[
\text{[Incidentul]}_\text{Event}_E \text{ a apărut } [\text{după} \text{ o dispută între individ şi personal} ]_\text{Time/Cause}_E \text{ [la o filială a Băncii Irlandeze din Cahir]}_\text{Place}_E.
\]

\section*{5.3 Results Analysis}

The initial experiment of importing semantic roles from one language to another has involved the translation of approx. 1000 sentences from the English FN.
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

The translations have been realized by professional translators. The reported problems during the translation relate mainly to the lack of the context of English sentences, which generate different translation variants. However, if the English semantic frame is considered, this problem is surmountable. The interface of the program offers the possibility to translate completely manually, or using as start-up the translation obtained with Google Translation service.

The alignment process was performed initially using the aligner developed by the Institute of Research in Artificial Intelligence, which is considered to have a precision of 87.17% and a recall of 70.25% (Tufiş et al., 2005). The import program offers now the possibility to align each English word with the Romanian translation manually, or by validating the alignment performed with GIZA++. It is recommended that the results are manually validated before entering the annotation import program.

The assessment of the correctness of the obtained Romanian corpus is performed by comparing it to the validated semantic role frames for Romanian (Trandabăţ and Husarciuc, 2008). The evaluation supposes comparing each English word’s semantic role annotation with the Romanian translation equivalent’s role. First results of the annotation import show an overall accuracy of approx. 79%. The validation focuses on detecting the cases where the import has failed, trying to discover if the problems are due to the translation/alignment phase or to the semantic or syntactic specificities of Romanian. Only few translation errors were found, and even then, the meaning has been kept and the semantic roles were correctly assigned. However, there are some mismatches, whose causes, partially analyzed in (Husarciuc et al., 2005), are (1) the double annotation, (2) the existence of imbricate frame elements (FEs) or (3) the unexpressed semantic frames.

5.3.1 Double Annotation

In the English FrameNet, a FE is double annotated if and only if, due to semantic ambiguity, its role in the sentence cannot be precisely established. The double annotation applies only to the non-core frame elements, due to the fact that the same phrase can refer to multiple circumstances (peripheral roles) of an event. When a semantic element is double annotated in English, the same generally
holds also for Romanian. The most frequent case of double annotation is for the Time/Cause roles, since many temporal specifications involve a cause and/or a goal. In the example:

En: Traditional methods require that [the animal]$_{Prot}$ [bleed]$_{Cause}$ to [death]$_{TARGET}$ [after having its throat cut]$_{Time/Cause}$; an agonizing and unsavory procedure lasting between one and two minutes.

Ro: Metodele tradiționale cer ca [animalul]$_{prot}$ [să sângereze]$_{cause}$ până la moarte$_{TARGET}$ [după ce i s-a tăiat gâtul]$_{Time/Cause}$; o procedură agonizantă și dezagreabilă durând între unul și două minute.

where the prepositional phrase is annotated both with a temporal and a causal semantic role, we have a succession of cause – effect relations: Cause$_1$ (cutting the throat) $\Rightarrow$ Effect$_1$ = Cause$_2$ (bleeding) $\Rightarrow$ Effect$_2$ (death).

The manual validation will resolve the cases in which a FE double annotated in English must be simple annotated in Romanian, if any. Until now, all the revised double annotation cases were cases where the ambiguity was kept in Romanian.

5.3.2 Imbrications

There are cases when a word can be part of two semantic elements without being double annotated. The imbrications process is common in the English annotations mainly in the possessive noun phrases (e.g. [[his]$_{Prot}$ ankle]$_{BodyPart}$).

In Romanian, when the possessive pronoun is placed before the verb as a reflexive pronoun, the imbrications disappear.

En: [When she got over the stroke]$_{Time/Cause}$ [she]$_{Exp}$ fell and [broke]$_{TARGET}$ [[her]$_{Exp}$ hand]$_{BodyPart}$.

Ro: [Când și-a revenit după atac]$_{Time/Cause}$, a căzut și [și]$_{Exp}$ - [a rupt]$_{TARGET}$ [mână]$_{BodyPart}$.

Even if we don’t have an absolute correspondence between the whole role BodyPart from English into Romanian, the noun mână (hand) is correctly annotated in Romanian as representing the BodyPart frame. The import of the annotation works also when the Romanian target-word is a gerund followed by a reflexive pronoun and a noun phrase, as in the following example:

En: [Josef Jakobs]$_{Prot}$ landed in a potato field in North Stifford, Essex, falling heavily and [breaking]$_{TARGET}$ [[his]$_{Prot}$ ankle]$_{BodyP}$. 

135
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

Ro: [Josef Jakobs]_{Prot} a aterizat într-un câmp de cartofi în North Stifford, Essex, căzând rău și [rupându]-{șî}_[TARGET-]_{Prot} gleznă_{BodyP}.

Although apparently similar to the English structure, in the Romanian sen-
tence, the frame elements are not imbricated, but successive, since the regent of
the pronoun șî is not the noun gleznă (ankle), but the gerundive verb.

5.3.3 Unexpressed Semantic Frames

A semantic role can be expressed in English, but implicit in Romanian, or vice-
versa. If the first case poses no problems to the transfer (a), the second one
supposes importing roles unexpressed in English (b).

(a) En: [Blood]_{Undergoer} [had congealed]_{TARGET} [thickly]_{Manner} [on the end of the smashed fibula]_{Place}.

Ro: [Sângele]_{Undergoer} [se îngroșă]_{TARGET} [spre capătul fibulei zdrobite]_{Place}.

(b) En: [Quit]_{TARGET} [smoking]_{Process}.

Ro: [Lăsatǎi]_{TARGET-} [vǎ]_{Protagonist} [de fumat]_{Process}.

In example (b) above, the English verb to quit is translated by a se lăsa, where the reflexive pronoun vǎ expresses the person that makes and supports, in
the same time, the action, therefore being the protagonist of the action. A situa-
tion apart is represented by the cases where uneaten food, had been translated
by mâncare, because the adjective corresponding of uneaten has, in Romanian,
the same root with the noun mâncare (food) and his utilization in such a case
would be inappropriate.

En: Remember that [any traces of uneaten food]_{Und} will [decom-
pose]_{TARGET} [in your tank]_{Place} and foul the water.

Ro: Nu uita că [orice resturi de mâncare]_{Und} [se vor descompu-
ne]_{TARGET} [în bazinul tǎu]_{Place} șì vor murdǎri apa.

The absence of the adjective in the Romanian sentence is imposed not only
by the syntactical or morphological specificity of the language, but also by some
pragmatic reasons. Somehow, any food is uneaten (yet). So “uneaten food”
is strongly sensed as pleonasm by native Romanian speakers. The expression
refers to an insignificant quantity of food left, by antithesis with the food already
ingested, implicitly mentioned. The semantics of this frame element differently
5.4 Conclusions

This chapter has presented a method of developing a Romanian resource of semantic roles through the transfer of the English annotation. The import method was preferred to the ‘classical’ creation by hand of a manually annotated corpus because it is a time-saving method. We believe the import method can be used for any other language, if an aligner is provided. Alternatively, we investigate a method of transferring the semantic annotation using syntactic trees for the two languages, instead of word alignments.

Tufiş et al. (2006) have tested an import method for enriching the Romanian WordNet with valence frames, transferred from the Czech WordNet. These frames are attached to verbs and specify syntactic and semantic restrictions for the arguments of the predicate, denoting the meaning of a given synset. We believe that the semantic role resource created for the Romanian language can be generalized to extract usage patterns for different verbs, containing syntactic and semantic information, similar to the ones imported from the Czech WordNet, which could be used to continue the development of the Romanian WordNet and to validate the valence frames imported from the Czech WordNet.

Since Romanian tends to be a less-resourced language, Cristea and Tufiş (2002) proposed the creation of a consortium, in order to assure an organized framework of communication between linguists and computer scientist interested in the Romanian language, as well as offering a repository of resources and tools for Romanian. The aim of this consortium is to align the Romanian language to the other European languages, from a point of view of resources, processing tools and visibility on the map of the languages used in electronic media, as started with the CLARIN project Cristea and Pistol (2009). An overview of the resources and tool collected within this consortium is presented in Cristea and Forăscu (2006). Adhering to the consortium goal, I started to contribute, working at different research projects within the Natural Language Processing Group of the Faculty of Computer Science of Iasi, to the development of resources and tools for the Romanian language (Pistol et al., 2006; Trandabăţ et al., 2006). My thesis is a further step in this direction, offering the semantic role resource presented in this
5. ROMANIAN RESOURCE OF SEMANTIC ROLES

chapter to the NLP community studying the Romanian language.
Applications of Semantic Role Labeling Systems

A large part of the work done in NLP deals with exploring how different tools and resources can be used to improve performance on a task. The quality and usefulness of the resource certainly is a major factor for the success of the research, but equally so is the creativity with which these tools or resources are used. There usually is more than one way to employ these, and the approach chosen largely determines the outcome of the work (Cui et al., 2005; Kwon et al., 2004; Narayanan and Harabagiu, 2004).

6.1 Semantic Roles in Question Answering

Question answering (QA) is the task of automatically answering a question posed in natural language using a collection of natural language documents. The Natural Language processing Group of the Faculty of Computer Science participated in the Question Answering competitions (QA@Clef\(^1\)) since 2006. The various stages of the systems are presented in several papers (Iftene et al., 2006, 2007b, 2009; Puşcaşu et al., 2007).

The architecture of the UAIC system that participated in the QA@CLEF RespubliQA competition in 2009 is presented in Figure 6.1. The collection of document from which the answer is to be extracted was for the 2009 competition

\(^1\)QA@Clef web page: http://nlp.uned.es/clef-qa/
6. APPLICATIONS OF SEMANTIC ROLE LABELING SYSTEMS

the JRC-Acquis corpus, a collection of juridical documents in XML format, with each paragraph numbered. The official nature of the documents has forced them to be in a very well organized structure, thus no corpus cleaning was necessary.

Figure 6.1: Architecture of the UAIC QA system that participated in the QA@Clef competition

The Question Analysis step is mainly concerned with the identification of the type of the answer. In order to determine it, the module identifies the question focus, the question type and a set of relevant keywords. The question analyzer performs the following steps:

- NP-chunking and Named Entity extraction;
- Question focus identification (the most important word in the sentence, justifying for the answer type);
- Answer type identification;
- Question type identification: the question types used in RespubliQA were: factoid, definition, purpose, reason and procedure;
6.1 Semantic Roles in Question Answering

- Identification of the keywords in the sentence that, together with the NPs and named entities are to be used by the query generator.

The next module retrieves, for every question, the relevant snippets of text from the document collection, using Lucene\(^2\) indexing and search tools. Queries are created based on the question analysis. They consist of the sequences of keywords previously identified, which are modified using some of the Lucene operators, such as score boosting (the “\(^{∧}\)” operator, followed by a positive integer), fuzzy matching (the “\(^{∼}\)” operator, followed by a number greater than 0 but less than 1) and the “or” operator (symbolized by words between parentheses).

The index was created using the XML files in the JRC-Acquis corpus. We have created two indexes, one at paragraph level and one at document level. The paragraph index is more precise in terms of relevant text, and is preferred for snippet extraction. If however, the answer is not found in the paragraph index, the query is applied to the document index instead. Using the queries, the indexes and Lucene search engine, we extract a ranked list of snippets for every question as possible answer candidates.

For this year’s track, we used the Factoid question answer extraction presented in Iftene et al. (2009), even if more detailed with sub-types (person, organization, count, measure, temporal, etc) and built special modules in order to extract answers for Definition and Reason questions. Simple pattern matching methods using rules extracted from the development questions / answers set was used for the other question types (Purpose and Procedure).

Our algorithm for answer extraction is based on several heuristics, but for similar result, Lucene scores are also considered. Examples of the heuristics we used are:

- the paragraph contains the question focus;

- the paragraph contains (at least) some of the name entities (directly proportional with the number of these name entities);

- if the question answer type is Person, or Organization, etc., we try to identify these types of named entities in the extracted paragraphs (and increase the Lucene score accordingly with the number of identified named entities);

\(^2\)Lucene web page: http://lucene.apache.org/
6. APPLICATIONS OF SEMANTIC ROLE LABELING SYSTEMS

- if the question type is Definition, then we prefer answers with definition form. In order to identify the definition forms, a grammar was built for identifying definitions in learning materials (Iftene et al., 2007a), and tailored to question-answering task as presented in (Iftene et al., 2008);

- the length of the sentence and the distance (in number of words) between the focus, the named entities and the keywords.

After applying these criteria, all paragraphs are awarded a score and the paragraph with the biggest score is chosen as containing the correct answer.

Evaluating the results of our QA system, we found that, from the 264 questions (out of 500) wrongly answered, only 35 had the correct answer in the ranked list returned by the retrieval module, and it was not correctly extracted. For the rest of them, the correct paragraph did not even entered in the candidate list. The question type that the module performed worst was the Purpose type.

One of the improvement methods considered was the use of the semantic role labeling system. Kaisser and Webber (2007) describe two new and complementary techniques for using semantic role resources and show the improvements to be gained when they are used individually or together. Compared with WordNet – which has been used widely in QA – FrameNet, PropBank and VerbNet are still relatively new. They offer the following features which can be used to gain a better understanding of questions, sentences containing answer candidates, and the relations between them:

- They all provide verb-argument structures for a large number of lexical entries.

- FrameNet and PropBank contain semantically annotated sentences that exemplify the underlying semantic frame.

- FrameNet (and PropBank through NomBank) contains not only verbs but also lexical entries for other part-of-speeches.

Kaisser and Webber (2007) have designed two methods that use these resources to annotate both questions and sentences containing answer candidates with semantic roles. If these annotations can successfully be matched, an answer candidate can be extracted. Thus, giving a complete frame semantic analysis of
the following sentences helps recognizing that they all contain an answer to the question *When was Alaska purchased*?:

(1) The United States purchased Alaska in 1867.

(2) Alaska was bought from Russia in 1867.

(3) In 1867, Russia sold Alaska to the United States.

(4) The acquisition of Alaska by the United States in 1867 is known as ‘Seward’s Folly’.

An algorithm that uses the three lexical resources to generate potential answer containing templates is presented in (Kaisser and Webber, 2007). While the templates contain holes – in particular, for the answer – the parts that are known can be used to create exact quoted search queries. Sentences can then be extracted from the output of the search engine and annotated with respect to the resource being used. From this, an answer candidate (if present) can be extracted. A second algorithm analyzes the dependency structure of the annotated example sentences in FrameNet and PropBank. It then poses rather abstract queries to the web, but can in its candidate sentence analysis stage deal with a wider range of syntactic possibilities. The two algorithms are complementary.

We considered using the first algorithm for the answer extraction step of our QA system. The first step is to annotate the question with its semantic roles. For this task, the best configuration discovered by PASRL was used. There are differences arising from the fact that the QA task requires to annotate questions, not declarative sentences:

1. The role labeler is trained on PropBank data, i.e. mostly on declarative sentences, whose syntax often differs considerably from the syntax of questions. As a result, the training and test set differ substantially in nature;

2. Questions tend to be shorter and simpler syntactically than declarative sentences - especially those occurring in news corpora;

3. Questions contain one semantic role that has to be annotated but which is not or is only implicitly (through the question word) mentioned – the answer.
The semantic type of the wanted answer is found and the snippets extracted by Lucene are reordered using as constraint if they contain the expected semantic role for the predicate of the question or not. For example, for the question:

(5) What has forced the simplification of the custom procedures?

the semantic role labeling system identifies the wh-word as the R-Arg0, indication that an agent is searched for in the document collection in a pattern such as:

ARG0 forced the simplification of the custom procedures.

The most frequent error of the semantic role labeling system, when applied to questions, was due to the different word order. Thus, for the sentence

(6) The creation of the council directive on the education of the migrating workers’ children aim at improving the conditions of freedom of movement.

the system correctly finds the predicate aim with its agent (Arg0) The creation of the council directive on the education of the migrating workers’ children, and the target (Arg2) at improving the conditions of freedom of movement, but when the text is in the form of the question:

(7) What does the creation of the council directive on the education of the migrating workers’ children aim at?

the system becomes confused by the long distance between the verb and the Wh-word that represent the target (since the semantic role labeling system takes into consideration words that are within a 3-words window around the predicate in the surface representation of the sentence).

Even with this drawback, the semantic role labeling system was successfully used by allowing Lucene to return a bigger number of paragraphs, which were annotated with semantic roles and ranked considering the fact if they contain or not the searched semantic role type.

6.2 Semantic Frames for Prosody Generation

Our investigation (Curteanu et al., 2009b) is based on fundamental studies and results in the last two decades, which established a reliable association between
6.2 Semantic Frames for Prosody Generation

intonational phrasing and meaning of the information structure (IS) notions, viz. Background-Kontrast entities (alias Topic-Focus in the Prague School’s language) and Theme-Rheme structures. Papers such as (Calhoun, 2003; Lee, 2007) support and prove, including at experimental and perception level, consistent rules that assign categorical functions of tones and tunes (boundary tones, pitch accents, and contours) to the IS textual entities and spans. Our main concern is on the textual IS-semantics side of the language interface within the (Romanian) prosody (Curteanu et al., 2007b), with the aim of squeezing all the discursive text meanings which proved to be useful for predicting a human-like prosody.

In (Curteanu et al., 2009a) two lines of investigation were followed for the intonational focus assignment in the Romanian sentence: (A) The Prague School’s Topic-Focus Articulation (TFA) algorithm is improved at clause level with hints from Van Valin’s linking algorithms(s) Van Valin (2005) and Gussenhoven’s SAAR (Sentence Accent Assignment Rule), then extended to inter-clause level. The information-structural (IS) textual spans of Theme(s)-Rheme(s) are computed within TFA approach as the lowest-highest degrees of Communicative Dynamism (CD) or predicative constituent actual ordering vs. the constituent Systemic Ordering (SO). (B) The second approach for computing Theme-Rheme and Background-Kontrast (or Topic-Focus in the Prague School’s language) is based on IS-semantics discourse theories, actually Asher’s SDRT (Segmented Discourse Representation Theory), and Heusinger’s attempt to design an Information Structure Discourse Theory (ISDT) (von Heusinger, 2004), which is embedded into SDRT. While Background-Kontrast entities are derived from SDRT on the basis of clause-level Given-New entity criteria, Theme-Rheme structures are computed with a recursive form of the Leong’s Inference-Boundary (IB) algorithm, applied for the first time to the Romanian sentence in (Curteanu et al., 2007c). The results of the two approaches (A) and (B) are distinct at Background-Kontrast and Theme-Rheme levels, providing different sets of ToBI annotations assigned to as Steedman’s rules.

Prague’s School Topic-Focus Articulation (TFA) algorithm (Hajicova et al., 1995) attributes Topic-Focus to a sentence, considering its constituent order in communicative dynamism (CD - the actual order of the constituents in speech) vs. the “standard” systemic ordering (SO - the statistically determined most frequent order of the constituents in speech), in the framework of contextual
boundness or non-boundness. The TFA algorithm applies on constituency parsed simple clauses, annotated with part-of-speech information and several semantic features, namely the specificity degrees (general - low specificity, contextually non-bound; specific - high specificity, contextually non-bound; indexical - mid-specificity, contextually-bound) of verbs and temporal / locative complements. The output of the TFA algorithm is determining the appurtenance of an element to the Topic or Focus elements of a sentence. In (Curteanu et al., 2007c), the TFA algorithm was adapted and implemented for the first time for Romanian prosodic structures. Examples 1 and 2 present sentences annotated with the TFA algorithm, where T stands for Topic, F for Focus, and T/F for ambiguous cases, and each syntactic constituent is represented within brackets:

(8) [Ion]_T [a câștigat]_F [competiția]_T/F.
John won the competition.

(9) [Privit de sus]_T, [licidul]_T [părea negru]_F.
From above, the liquid seemed black.

Applying the TFA algorithm on Romanian raised a number of issues. First, the context considered for the boundness constraints is minimal, addressing only the current sentence (whether the head of a noun group is defined or not, or whether the verb or the complementation are indexical references), although Hajicova et al. (1995) mentioned that the previous sentence is needed in order to properly analyze a verb context. Another problem is due to the fact that, following the Topic-Focus assignment, some sentences contain no focus at all, or more than one focused constituent. We need a balanced assignment of Topic-Focus entities, observing the TFA criteria, intonational grouping, and sentence level prosody patterns such as Sentence Accent Assignment Rule (SAAR) (Gussenhoven, 2007).

We propose (Curteanu et al., 2009b) a slightly extended version of the original TFA algorithm (Hajicova et al., 1995), where the syntactic and semantic informa-

---

3Examples of specificity degrees for temporal complementation are: general - niciodată, mereu, indexical - astăzi, anul acesta, specific - 22 iunie, într-o frumosă zi de mai.

4“As for the verb, it is important to have access to the verb of the preceding utterance and to use a systematic semantic classification of the verbs. If the main verb of sentence n has the same meaning as (or a meaning included in) that of sentence n – 1 (in the sense of hyponymy), then it belongs to the topic.”
tion required by the original algorithm is completed with IS annotation, obtained by discursive analysis. The starting point of our approach is the fact that the TFA notion of Topic has much in common with the more recently characterized concept of Background, while the Focus correspond to the notion of Kontrast (Lee, 2007). Thus, considering the preceding context of a sentence (the sentences already uttered in the discourse, i.e. the accessibility domain), and performing co-reference resolution, the contextually-boundness information is obtained using anaphoric relators. Each constituent is thus annotated with Background or Kontrast, where the backgrounded entities are the ones already mentioned in the discourse (represented in many cases by definite noun groups, indexical verbs or complements, personal pronouns, but not only this, as considered by the original TFA algorithm), and kontrasted entities are those that have not been previously introduced in the discourse (in the original TFA, they correspond to undefined noun groups, specific verbs or complements).

Hajicova et al. (1995) states that “In its present form, however, the algorithm has several limitations. It can process only simple sentences.”. Thus, a further step in the development of the TFA algorithm for Romanian was the extension of the TFA algorithm to complex sentences. A simple and effective method is splitting the complex clause into simple, finite clauses, and then applying the algorithm recursively on each clause. In correlated adjacent clauses, the algorithm is simply applied consecutively for each clause. More attention needs to be assessed in applying the algorithm on subordinate clauses, since the subordinate clause is to be treated as the corresponding complement of the verb, completing thus the regent sentence. For example, the complex clause in example 3 contains two simple clauses, marked by /1 and /2.

(10) \([\text{Ion}]_T [\text{\c{s}tia}]_F^{/1} \text{\c{s}a}\ [\text{\int\c{a}rziase}]_F [\text{mult}]_T^{/2}.
\text{John} \text{ knew } \text{that he was } \text{very late.}

This example can be transformed into a simple clause, by reducing the completive clause to a direct argument:

(11) \([\text{Ion}]_T [\text{\c{s}tia}]_F \text{asta}_{/F}.
\text{John} \text{ knew that.}

This compression of the complex clauses into mere complements allows for the
6. APPLICATIONS OF SEMANTIC ROLE LABELING SYSTEMS

appliance of the TFA algorithm on complex sentences. The benefits of this approach become evident if more than one subordinate clause is considered, as in examples (12) and (13).

(12)  [Ion] [a anunțat]/\(^{1a}\) [ceea ce a descoperit]/\(^{2}\)
       John announced what he has discovered
       [în fața întregii audienței]/\(^{1b}\)
       in front of the audience.

The complex clause in example (12) is formed by two simple clauses, but one of them is intercalated in between two others (the clause “what he has discovered” is embedded into the second clause “John announced * in front of the audience”). If we consider the embedded completive clause a simple complement, the TFA algorithm will yield the result presented in (13).

(13)  [Ion] \(_T\) [a anunțat]/\(_T/F\) [ceva]/\(_T/F\) [în fața întregii audienței]/\(_F\)
       John announced something in front of the audience.

The examples above show that the Background-Kontrast extended version of the TFA algorithm works on both simple clauses and complex sentences, providing Background-Kontrast (or Topic-Focus) annotation that can be transformed into pitch accents by using the rules described in Steedman (2000).

After performing Background-Kontrast (Topic-Focus) assignment on the sentence constituents using the extended TFA presented above, sometimes more than one focus appear, or a sentence may contain no focus at all, only ambiguous T/F’s. We consider that in a sentence may occur several focused constituents, but only if they belong to different intonational groups. In fact, analyzing the manually annotated data, we found that each intonational group (i.e. intermediate phrase, \(_{ip}^{5}\)) must contain one (and only one) focus (kontrast), the rest of the elements in the intonational group being topics (backgrounds). In order to assure that each intermediate phrase will receive an unique focus tag, we developed an algorithm that filters if more than one focus appear in an intermediate phrase, or transforms a T/F or T marking to F if no focus is found in the intermediate phrase, using a set of rules derived from Gussenhoven’s Sentence Accent Assign-

\(^{5}_{ip}\) (intermediate phrase) - each clause is an \(_{ip}\). There are cases when \(_{ip}\) structures are interrupted by speech break markers (punctuation marks, coordinative conjunctions, etc.); \(_{IP}\) (Intonational Phrase) - each sentence, together with all its subordinate clauses, forms an \(_{IP}\).
6.2 Semantic Frames for Prosody Generation

The rules presented below apply recursively for all the intonational units in the sentence. In the algorithm presented below F, T and T/F stand for three sets, each containing the constituents from the intonational phrase that are marked with the respective tags after applying the extended TFA algorithm.

\( iF = \emptyset \)

/*none of the sentence constituents received focus and the focus set is empty*/
apply SAAR-derived rules to decide which constituent turns to F from the T or T/F sets
The other t/f marked constituents are changed to T.

\( iF > 1 \)
/*two or more sentence constituents received focus*/
Apply SAAR-derived rule to decide which constituent keeps the focus;
The other F-marked constituents receive T instead of F.
The other T/F marked constituents are changed to T.

The SAAR-derived rules that we apply are:

1. If the verb has adjuncts, then the adjuncts will receive focus rather than the arguments.

2. If the verb has arguments, then the adjuncts will receive focus rather than the verb.

3. If the verb has more than one adjunct or more than one argument, then the rightmost one receives the focus.

The informal meaning of the SAAR-derived rules is that the semantic periphery is preferred for intonational focus (kontrast) rather than its semantic head, in the ascending intonational accentuation order head < sub F argument < sub F adjunct (for finite / non-finite clause) and head < sub F modifier (for nominal, adjectival, verbal groups). In what follows, several examples of refining the extended version of the TFA algorithm with intonational phrasing sequences are discussed:

\[
(14) \quad [\text{Ion}]_T \ [a \ c\breve{a}stigat]_F \ [\text{competi}t\breve{t}ia]_{T/F}.
\end{equation}
\[
\{[[\text{Ion}] \ [a \ c\breve{a}stigat] \ [\text{competi}t\breve{t}ia]]_{ip}\}_IP.
\]
6. APPLICATIONS OF SEMANTIC ROLE LABELING SYSTEMS

\[
\begin{align*}
\{\{[\text{Ion}]_T \text{ [a câştigat]}_F \text{ [competiţia]}_{IP}\}_p\}_IP.
\end{align*}
\]
John won the competition.

In this example, the last constituent changes from T/F to T since the intermediate phrase can have only one focused constituent.

(15) \[
[\text{Privit de sus}]_T, \quad [\text{lichidul}]_T \quad [\text{prea negru}]_F.
\]
\[
\{[\text{Privit de sus}]_{IP}\}, \{[[\text{lichidul}]_T \text{ [prea negru]}_T]_{ip}\}_IP.
\]
\[
\{[\text{Privit de sus}]_F\}_IP, \{[[\text{lichidul}]_T \text{ [prea negru]}_T]_{ip}\}_IP.
\]

From above, the liquid seemed black.

In example (15), the first ip has no focus, hence the constituent “from above”, being the only constituent in the intermediate phrase, receives F instead of T.

(16) \[
[\text{Maria}]_T \quad [\text{nu a fost}]_T \quad [\text{în acea zi}]_T
\]
\[
\{[[\text{Maria}]_T \text{ [nu a fost]}_T \text{ [în acea zi]}_T]_{ip}\}_IP.
\]
\[
\{[[\text{Maria}]_T \text{ [nu a fost]}_T \text{ [în acea zi]}_F\}_IP.
\]

Mary wasn’t that day.

In example (16), the intermediate phrase has no focus, and the verb has an argument (the Agent “Mary”) and an adjunct (the temporal “that day”). According to the defined SAAR-rules, the adjunct will receive focus.

(17) \[
[\text{Ion}]_T \quad [\text{a câştigat}]_F \quad [\text{o competiţie}]_F.
\]
\[
\{[[\text{Ion}]_T \text{ [a câştigat]}_T \text{ [o competiţie]}_F]_{ip}\}_IP.
\]
\[
\{[[\text{Ion}]_T \text{ [a câştigat]}_T \text{ [o competiţie]}_F]_{ip}\}_IP.
\]

In example (17), the intermediate phrase has two focused constituents; therefore the verb will lose its focus in favor of its argument “a competition”.

The examples above show that the extended TFA algorithm can be further improved, considering the intonational grouping of the sentence.

In order to assess the results of the presented algorithms, we compared the output with the manual annotation of the same sentences (a set of spoken Ro-
manian sentences extracted from George Orwell’s novel 1984). The sentences were marked with Theme-Rheme boundaries and with ToBI pitch accents by the Group of Speech Processing within the Institute of Computer Science. The results showed that assigning Background-Kontrast (Topic-Focus) considering the intonational phrasing and the SAAR-derived rules is a very good practice, especially for the cases where the extended TFA algorithm returned no focused constituents. The main differences observed concerned the marking of the last constituent of the sentence: in the TFA approach, it is mostly focused (being the rightmost constituent is one of the major reasons to focus it); in the gold annotation, the last constituent marks clearly the descending tendency of declarative sentences in Romanian, which are usually marked with a low tone.

We believe that the prediction of prosody from text can be substantially improved by a deeper understanding of the textual discourse theories, with the natural emphasis on the information structure (IS) discourse semantics, using also proper degrees of communicative dynamism for the surface representations of a sentence. We presented an extension of the classic TFA algorithm (Hajicova et al., 1995) for complex clauses, using also discourse co-referred entities besides syntactical information (nouns determined or not, indexical verbs or complements). The extended TFA uses SAAR-derived rules in order to refine the Background-Kontrast (Topic-Focus) assignment and to eliminate ambiguities.

6.3 Conclusions

This section presented two possible applications of the semantic roles in natural language processing. For a Question Answering system, semantic roles are very powerful in the snippet ranking and answer extraction phase. For spoken language and prosody generation, semantic roles are a good linker between the textual representation and the intonational accentuation.
6. APPLICATIONS OF SEMANTIC ROLE LABELING SYSTEMS
Conclusions and Further Work

Semantic roles are one of the major steps in representing text meaning, referring to finding the semantic relations between a predicate and syntactic constituents in a sentence. All content elements of a language are seen as predicates, i.e. expressions which designate events, properties of, or relations between, entities. Predication is the mechanism that allows individuals to instantiate properties, actions, attributes and states. Linguistic expressions can be dependent or independent. For example, the word hat can be understood outside any circumstance, time, or person, because it does not have to be attributed to anything or anyone: it is an individual. On the contrary, if we consider the word red, the denotations for this word cannot be understood outside its association with an individual: red hat. In linguistic terms, the dependent phenomena are predicates, while individuals are arguments. The linking between a phenomenon and individuals is known as predication. Predicates are not treated as isolated elements, but as structures, named semantic frames. Within the predicate frames, each entity plays a role, called semantic role. Semantic roles are semantic relations that connect entities to events, more particularly, arguments to predicates. Chapter 2 of the thesis presents the linguistics behind the semantic role theories, types of semantic roles and their characteristics. The intuition that semantic analysis can make a positive contribution to language-based applications has motivated the development of a number of lexical-semantic resources. Prominent among them are PropBank and FrameNet. The potential contribution of these resources is constrained by the information they contain and the level of effort involved in their development.
Some critics have been raised, mainly due to the fact that there has never been agreement on a (small) fixed set of thematic roles, and every resource uses its own technique. Another frequent point of difference, for example, is whether to treat Recipient and Experiencer as distinct thematic roles or to lump one or both with goal. The requirement that exactly one role be assigned to each argument of a verb has been attacked as well, notably by Lakoff (1986).

Recognizing and labeling semantic arguments is a key task for answering Who did what, when, where, why, etc. The importance of the task is confirmed by the emergence of international competitions concerned with the recognition of semantic roles (mainly for the English language), based on PropBank predicate-argument structures (Shared Tasks of CoNLL-2005-2009) or FrameNet semantic frames (SemEval 2007). Given a sentence, the task consists of analyzing the scene expressed by some predicational word of the sentence. In particular, for each predicational verb or noun (referred as target word), all the constituents in the sentence which fill a semantic role of the target word have to be recognized and labeled accordingly. This problem has been referred as Semantic Role Labeling (SRL) and the most common methods to perform SRL have been presented in Chapter 3. The work on SRL has included a broad spectrum of probabilistic and machine-learning approaches to the task, mostly supervised systems, because most SRL research takes an approach requiring training on role-annotated data. The existing studies have used different statistical frameworks, but have largely converged on a common set of features to base their decisions on, namely syntactic information (path from predicate to constituent, phrasal type of constituent) and lexical information (head word of the constituent, predicate).

This thesis introduces two Semantic Role Labeling systems developed for English to annotate raw text with the semantic role information, for usage in natural language processing tools: one rule-based, baseline system, and a semantic role labeler based on extensive use of machine learning algorithms. The rule-based system uses a set of rules discovered by analysis the English sentences annotated with semantic roles, but has limited performances (63.79%). The second system uses as training set a collection of sentences with annotated semantic roles. The system trains different machine learning algorithms using the framework provided by Weka (Witten and Frank, 2005), evaluates their results using 10-fold cross-validation, and saves the best model as part of the Semantic Role Labeling
system. The systems presented in Chapter 4 are intended to annotate new, raw texts, performing pre-processing to add part-of-speech and syntactic dependencies, then dividing the task of Semantic Role Labeling into different tasks:

**Predicate Identification** – this module takes the syntactic analyzed sentence and decides which of its verbs and nouns are predicational, thus whom semantic roles need to be searched for;

**Predicate Sense Identification** – once the predicates for a sentence are marked, each predicate need to be disambiguated in order to select its appropriate role set;

**Sense Identification** the Predicate and Predicate Sense Identification tasks are simultaneously performed;

**Semantic Roles Identification** – identify a semantic role for each of the syntactic dependents of selected predicates.

The performances of the second system are similar to the state-of-the art SRL systems for English (85%), and even better when the Czech or Chinese corpus are considered as training sets.

The main question this thesis intends to answer to is if semantic role information is cross-linguistically valid, and if so, up to what extent. The interest begun when observing the huge amount of time and human resources involved in creating the semantic role resources for English and in later years for German, Spanish and Japanese also. Since semantic information is considered of major influence for a natural language processing system, we started to consider developing such a resource for Romanian, but with considerable less human and temporal resources. Thus, we investigated the transfer of semantic role annotation from one language to another (from English to Romanian in our case) and presented in Chapter 5 a semantic role annotation transfer method.

However, semantic frames continue to be an exciting research subject, fact evidenced also by the large number of NLP applications that considers it. In the context of the recent Question Answering challenge, frame-semantic representations have been applied successfully to approximate information retrieval patterns.

The main contributions of this thesis are:
7. CONCLUSIONS AND FURTHER WORK

- Introducing a method to create a semantic role annotated corpus with minimum resources, through the transfer of the annotation from English to another language. The method is applied to Romanian, and we believe it can be successfully used for other language;

- Creating a resource of annotated Semantic Roles for Romanian by using the proposed method - the Semantic Role resource for Romanian is available at http://students.info.uaic.ro/~dtrandabat/pages/phdThesis.php;

- Providing an interface that can be used to easily further develop the semantic role resource;

- Developing a rule-based semantic role labeling system for English;

- Creating a platform for developing adjustable supervised semantic role labelers. This platform trains different machine learning algorithms on a training set, selects the best performing algorithm and builds a semantic role labeler using the best models;

- Testing the platform by creating a Semantic Role Labeler for English and Romanian;

- Contributing to NLP applications by integrating semantic frames in a Question - Answering system and by designing the improvement strategy of a Prosody - generation system for synthesized speech based on semantic roles.

The semantic roles encode important information that can improve natural language processing systems. Therefore, beside the applications presented in Chapter 6, based on the experience of the Natural Language Processing Group of the Faculty of Computer Science, several other interactions between semantic roles and different tasks of computational linguistics can be envisaged. The most immediate application is the use of semantic roles for anaphora resolution, approaches in this sense existing already (Ponzetto and Strube, 2006). The experience in anaphora resolution (Cristea and Postolache, 2005) showed that the pronominal anaphoric relations are sometimes hard to identify. As an example, consider The decision was adopted by the council; it published it. Identifying the semantic roles for the first sentence (the decision as Patient
and the council as Agent), as well as for the second sentence (the first it is the Agent, and the second one is a Patient), and using the semantic parallelism presented in Mitkov (1999), which states that phrases that have the same semantic role as the anaphor, are favored, we can easily identify that the Agent of the first sentence, the council, corresponds to the agent of the second sentence, the first it.

A field in which the semantic roles may be equally of help is the detection of subjectivity in texts. In Trandabăț et al. (2007), the influence of the syntactic inner structure of the verbal group was investigated in relation to subjectivity, and the result showed that different sentiment may trigger different constructions. Transferring the analysis to semantic relations seems an interesting approach.

Another envisaged application of the semantic roles is in summarization. Cristea et al. (2005) present a method of realizing summaries based on the vein structure (Cristea et al., 1998; Ide and Cristea, 2000) of the discourse. The semantic roles may be used to refine the vein structure of the discourse, and thus improving the summaries. To give just an example, identifying the adjuncts can definitely help establishing what information can be omitted from a summary.

In conclusion, we consider that the thesis, providing a semantic role resource for Romanian, semantic role labeling systems for English and Romanian, and a platform for the development of semantic role annotation systems for other languages, represents a solid starting point in introducing semantic processing to Romanian NLP systems.
7. CONCLUSIONS AND FURTHER WORK
References


Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie, and Angela D. Friederici. Semantic role universals and argument linking: theoretical, typological, and psycholinguistic perspectives. Mouton de Gruyter, 2006. 11


REFERENCES


Romanian Academy, Institute of Romanian Philology ”A. Philippide”, Trinitas Publishing House, Iaşi, Romania, 2002. 137


N. Curteanu. Contrastive meanings of the terms ”predicative” and ”predicational” in various linguistic theories (i, ii). Computer Science Journal of Moldova (R. Moldova), 11(4), 2003. 15

Neculai Curteanu and Diana Trandabăt. Functional (F)X bar projections for local and global text structures. the anatomy of predication. Revue Roumaine de Linguistique, 50, January–July 2007. ISSN 0035-3957. 18


REFERENCES

Neculai Curteanu, Diana Trandabăt, and Mihai-Alex Moruz. Functional FX bar projections of the (Romanian) verbal group and sub-groups on the syntactic-semantic interface. Computer Science Journal of Moldova (R. Moldova), 15:123–152, 2007a. ISSN 1561-4042. 16, 18


REFERENCES


164
REFERENCES


REFERENCES


Namhee Kwon, Michael Fleischman, and Eduard Hovy. SENSEVAL automatic labeling of semantic roles using maximum entropy models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, SENSEVAL-3 Third
REFERENCES


Monica Lupu, Diana Trandabăţ, and Maria Husarciuc. A Romanian SemCor aligned to the English and Italian MultiSemCor. In 1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School, Proceedings, pages 20–27, Cluj-Napoca, Romania, July 2005. 120


REFERENCES


REFERENCES


Ferdinand de Saussure. *Cours de Linguistique générale*. Payot, 1916. 1, 52


REFERENCES


Diana Trandabăț, Maria Husarciuc, and Monica Lupu. Towards an automatic import of English FrameNet frames into the Romanian language. In ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School, pages 28–36, Cluj-Napoca, Romania, July 2005. 120


REFERENCES

Appendices

Appendix A

PropBank Annotation Example for the verb take

<frameset>
<note>Frames file for ‘take’ based on survey of sentences from big corpus</note>
<predicate lemma="take">
<roleset id="take.01" name="take, acquire, come to have"
  vncls="10.5 11.3">
<roles>
<role descr="Taker" n="0">  
<vnrole vncls="10.5" vntheta="Agent" />
<vnrole vncls="11.3" vntheta="Agent" />
</role>
<role descr="thing taken" n="1">  
<vnrole vncls="10.5" vntheta="Theme" />
<vnrole vncls="11.3" vntheta="Theme" />
</role>
<role descr="entity taken from, or prepositional complement of arg1" n="2">  
<vnrole vncls="10.5" vntheta="Source" />
<vnrole vncls="11.3" vntheta="Source" />
</role>
</roles>
<example name="take office">  
<text>President Carlos Menem took office July 8.</text>
<arg n="0">President Carlos Menem</arg>
<rel>took</rel>
<arg n="1">office</arg>
<arg f="TMP" n="M">July 8</arg>
</example>
<example name="extract">  
<text>Sony took a lesson from the American management books.</text>
<arg n="0">Sony</arg>
<rel>took</rel>
<arg n="1">a lesson</arg>
<arg f="from" n="2">the American management books</arg>
</example>
</roleset>
</predicate>
</frameset>
Principals take cheating seriously

She took the law into her own hands

It has taken measures to prevent cheating.

According to industry lawyers, the ruling gives pipeline companies an important second chance to resolve remaining disputes and take advantage of the cost-sharing mechanism.

Investors unsettled by the stock market’s recent gyrations can take some comfort in the predictable arrival of quarterly dividend checks.

Frank plans the program, takes care of business, and approaches the work like any other job.
<arg n="0">Frank</arg> <rel>takes</rel> <arg n="1">care</arg> <arg n="2" f="of">business</arg> </example> </roleset> <roleset id="take.02" name="tolerate" vncls="-"> <roles> <role descr="tolerator" n="0" /> <role descr="thing tolerated" n="1" /> </roles> <example name="put up or shut up"> <text>John couldn’t take his medicine like a man.</text> <arg n="0">John</arg> <arg f="MOD" n="m">could</arg> <rel>take</rel> <arg n="1">his medicine</arg> <arg f="MNR" n="m">like a man</arg> </example> </roleset> <roleset id="take.03" name="cause (to be)" vncls="-"> <roles> <role descr="impeller to action" n="0" /> <role descr="impelled agent, topic" n="1" /> <role descr="impelled action, attribute of arg1" n="2" /> </roles> <example name="resultative"> <text>John took his company private.</text> <arg n="0">John</arg> <rel>took</rel> <arg n="1">his company</arg> <arg n="2">private</arg> </example> </roleset> <roleset id="take.04" name="understand to be" vncls="29.2"> <roles> <role descr="understander" n="0"> <vnrole vncls="29.2" vnttheta="Agent" /> </role> <role descr="thing" n="1"> <vnrole vncls="29.2" vnttheta="Theme" /> </role> </roleset>
I take it, then, that John is a complete and utter idiot?

Note: Again, this is a 'dummy' it, without any real semantic substance.

These events took place 35 years ago.

John's company took a $5 million write-off due to unwise investments in jellybeans.

Note: This is a 'dummy' need, without any real semantic substance.
Being so fragile and minute, they will take special special robotic handling equipment.

It took me a half-hour to move 10 feet from my parking spot in an outer lot to an aisle, and an additional hour to reach an inner roadway a half-block away.

Rumors of an impending devaluation have been circulating in Moscow for weeks, but the size of the cut took many Western bankers by surprise.
<example name="ballooning? really?"><inflection person="ns" tense="present" aspect="perfect" voice="active" form="participle" /></example>
<text>Americans it seems [0] [∗Tx−1] have followed Malcolm Forbes 's hot-air lead and taken to ballooning in a heady way.</text>
<arg n="0">Americans</arg>
<arg n="M" f="DIS">it seems [0] [∗Tx−1]</arg>
<rel>taken</rel>
<arg n="1" f="to">ballooning</arg>
<arg n="M" f="MNR">in a heady way</arg>
</example>
</roleset>
</predicate>

<predicate lemma="take_away">
<roleset id="take.05" name="take away: remove" vncls="10.5">
<roles>
<role descr="taker" n="0">vrole vncls="10.5" vtheta="Agent" /></role>
<role descr="thing removed" n="1">vrole vncls="10.5" vtheta="Theme" /></role>
</roles>
</roleset>
</predicate>

<predicate lemma="take_in">
<roleset id="take.06" name="take in: give shelter" vncls="-">
<roles>
<role descr="entity giving shelter" n="0" /></role>
<role descr="entity taking shelter" n="1" /></role>
</roles>
</roleset>
</predicate>

<example name="Calgon , take me away!"> Marshall Coleman wants to *trace* take away your right to choose </example>
<arg n="0">*trace* → Marshall Coleman</arg>
<rel>take away</rel>
<arg n="1">your right to choose</arg>
</example>
</roleset>
</predicate>

<example name="Arg0 rel Arg1">
<text>This exclusive club has taken in a host of flashy new members</text>
<arg n="0">This exclusive club</arg>
</example>
</roleset>
</predicate>
<rel>taken in</rel>
<arg n="1">a host of flashy new members</arg>
</example>
</roleset>
</predicate>

<predicate lemma="take_off">
<roleset id="take.07" name="take off: remove" vncls="10.5">
<roles>
  <role descr="entity removing something" n="0">"/>
  <vnrole vncls="10.5" vnttheta="Agent"/>
</role>
  <role descr="thing being removed" n="1">"/>
  <vnrole vncls="10.5" vnttheta="Theme"/>
</role>
</roleset>
<example name="take it all off">
<text>John took off all his clothes in public. </text>
<arg n="0">John</arg>
<rel>took off</rel>
<arg n="1">all his clothes</arg>
<arg f="LOC" n="M">in public</arg>
</example>
</roleset>
<roleset id="take.08" name="take off: increase dramatically" vncls="-">
<roles>
  <role descr="thing increasing" n="1"/>
</roleset>
<example name="zoom">
<text>The pace has taken off. </text>
<arg n="1">The pace</arg>
<rel>taken off</rel>
</example>
</roleset>
<roleset id="take.19" vncls="-" name="take off: like a plane">
<roles>
  <role n="1" descr="airplane"/>
  <role n="2" descr="airport"/>
</roleset>
<example name="leavin' on a jet plane...">
<inflection person="third" tense="present" aspect="ns" voice="active" form="full"/>
<text>Keep in mind that this is the same movie in which a character is flattened by a steamroller only to pop right back up and peer in the window of a Boeing 747 — from the outside — as it takes off. </text>
<arg n="1">it</arg>
<rel>[ takes | off]"/>
And when Mr. Engelken asked him why he took time off from work for somebody he didn’t even know...

He says with a shake of the head.

'I will sit down and talk some of the problems out, but take on the political system? Uh-uh,' he says with a shake of the head.
I will take on the political system.

John took over the company.

John took over as president.

John took up as president.
Most of the picture is taken up with endless scenes of talking heads.

Most of the picture

taken up

with endless scenes of talking heads

The Senate will take up the measure quickly.

The Senate

take up

the measure

quickly

They took it up with Warner.

They

took up

it

Warner

They are taken back in an interview when asked whether, as mayor, he plans on reforming the political ‘‘fiefdoms’’ that perpetuate the monumental ineffectiveness of New York’s school system.
Appendix B

WordNet senses for the verb take

1. take (carry out) “take action”; “take steps”; “take vengeance”.
2. take, occupy, use up (require (time or space)) “It took three hours to get to work this morning”; “This event occupied a very short time”.
3. lead, take, direct, conduct, guide (take somebody somewhere) “We lead him to our chief”; “can you take me to the main entrance?”; “He conducted us to the palace”.
4. take, get hold of (get into one’s hands, take physically) “Take a cookie!”; “Can you take this bag, please”.
5. assume, acquire, adopt, take on, take (take on a certain form, attribute, or aspect) “His voice took on a sad tone”; “The story took a new turn”; “he adopted an air of superiority”; “She assumed strange manners”; “The gods assume human or animal form in these fables”.
6. take, read (interpret something in a certain way; convey a particular meaning or impression) “I read this address as a satire”; “How should I take this message?”; “You can’t take credit for this!”.
7. bring, convey, take (take something or somebody with oneself somewhere) “Bring me the box from the other room”; “Take these letters to the boss”; “This brings me to the main point”.
8. take (take into one’s possession) “We are taking an orphan from Romania”; “I’ll take three salmon steaks”.
9. take (travel or go by means of a certain kind of transportation, or a certain route) “He takes the bus to work”; “She takes Route 1 to Newark”.
10. choose, take, select, pick out (pick out, select, or choose from a number of alternatives) “Take any one of these cards”; “Choose a good husband for your daughter”; “She selected a pair of shoes from among the dozen the salesgirl had shown her”.
11. accept, take, have (receive willingly something given or offered) “The only girl who would have him was the miller’s daughter”; “I won’t have this dog in my house!”; “Please accept my present”.

187
12. fill, take, occupy (assume, as of positions or roles) “She took the job as
director of development”; “he occupies the position of manager”; “the young
prince will soon occupy the throne”.
13. consider, take, deal, look at (take into consideration for exemplifying pur-
poses) “Take the case of China”; “Consider the following case”.
14. necessitate, ask, postulate, need, require, take, involve, call for, demand
(require as useful, just, or proper) “It takes nerve to do what she did”; “success usually requires hard work”; “This job asks a lot of patience and
skill”; “This position demands a lot of personal sacrifice”; “This dinner calls
for a spectacular dessert”; “This intervention does not postulate a patient’s
consent”.
15. take (experience or feel or submit to) “Take a test”; “Take the plunge”.
16. film, shoot, take (make a film or photograph of something) “take a scene”;
“shoot a movie”.
17. remove, take, take away, withdraw (remove something concrete, as by lifting,
pushing, or taking off, or remove something abstract) “remove a threat”; “remove a wrapper”; “Remove the dirty dishes from the table”; “take the
gun from your pocket”; “This machine withdraws heat from the environ-
ment”.
18. consume, ingest, take in, take, have (serve oneself to, or consume regularly)
“Have another bowl of chicken soup!”; “I don’t take sugar in my coffee”. 
19. take, submit (accept or undergo, often unwillingly) “We took a pay cut”.
20. take, accept (make use of or accept for some purpose) “take a risk”; “take
an opportunity”.
21. take (take by force) “Hitler took the Baltic Republics”; “The army took the
fort on the hill”.
22. assume, take, strike, take up (occupy or take on) “He assumes the lotus
position”; “She took her seat on the stage”; “We took our seats in the
orchestra”; “She took up her position behind the tree”; “strike a pose”.
23. accept, admit, take, take on (admit into a group or community) “accept
students for graduate study”; “We’ll have to vote on whether or not to
admit a new member”.
24. take (ascertain or determine by measuring, computing or take a reading
from a dial) “take a pulse”; “A reading was taken of the earth’s tremors”.
25. learn, study, read, take (be a student of a certain subject) “She is reading
for the bar exam”. 

188
26. claim, take, exact (take as an undesirable consequence of some event or state of affairs) “the accident claimed three lives”; “The hard work took its toll on her”.
27. take, make (head into a specified direction) “The escaped convict took to the hills”; “We made for the mountains”.
28. aim, take, train, take aim, direct (point or cause to go (blows, weapons, or objects such as photographic equipment) towards) “Please don’t aim at your little brother!”; “He trained his gun on the burglar”; “Don’t train your camera on the women”; “Take a swipe at one’s opponent”.
29. take (be seized or affected in a specified way) “take sick”; “be taken drunk”.
30. carry, pack, take (have with oneself; have on one’s person) “She always takes an umbrella”; “I always carry money”; “She packs a gun when she goes into the mountains”.
31. lease, rent, hire, charter, engage, take (engage for service under a term of contract) “We took an apartment on a quiet street”; “Let’s rent a car”; “Shall we take a guide in Rome?”.
32. subscribe, subscribe to, take (receive or obtain regularly) “We take the Times every day”.
33. take (buy, select) “I’ll take a pound of that sausage”.
34. take (to get into a position of having, e.g., safety, comfort) “take shelter from the storm”.
35. take, have (have sex with; archaic use) “He had taken this woman when she was most vulnerable”.
36. claim, take (lay claim to; as of an idea) “She took credit for the whole idea”.
37. accept, take (be designed to hold or take) “This surface will not take the dye”.
38. contain, take, hold (be capable of holding or containing) “This box won’t take all the items”; “The flask holds one gallon”.
39. take (develop a habit) “He took to visiting bars”.
40. drive, take (proceed along in a vehicle) “We drive the turnpike to work”.
41. take (obtain by winning) “Winner takes all”; “He took first prize”.
42. contract, take, get (be stricken by an illness, fall victim to an illness) “He got AIDS”; “She came down with pneumonia”; “She took a chill”.

Beside the senses for the verb take, WordNet lists senses also for the phrasal verbs in PropBank (take away, take off, take out, etc.)
and for two of the expressions in PropBank (take advantage and take care). These senses are presented below.

**take away**

1. take away, bear off, bear away, carry away, carry off – (remove from a certain place, environment, or mental or emotional state; transport into a new location or state; “Their dreams carried the Romantics away into distant lands”; “The car carried us off to the meeting”; “I'll take you away on a holiday”; “I got carried away when I saw the dead man and I started to cry”).

2. remove, take, take away, withdraw – (remove something concrete, as by lifting, pushing, or taking off, or remove something abstract; “remove a threat”; “remove a wrapper”; “Remove the dirty dishes from the table”; “take the gun from your pocket”; “This machine withdraws heat from the environment”).

3. take away, take out – (take out or remove; “take out the chicken after adding the vegetables”).

4. take away – (take from a person or place; “We took the abused child away from its parents”).

5. take out, take away – (buy and consume food from a restaurant or establishment that sells prepared food; “We’ll take out pizza, since I am too tired to cook”).

6. remove, take away – (get rid of something abstract; “The death of her mother removed the last obstacle to their marriage”; “God takes away your sins”).

7. take away, detract – (take away a part from; diminish; “His bad manners detract from his good character”).

**take in**

1. take in – (provide with shelter).

2. gull, dupe, slang,befool, cod, fool, put on, take in, put one over, put one across – (fool or hoax; “The immigrant was duped because he trusted everyone”; “You can’t fool me!”).

3. absorb, take in – (suck or take up or in; “A black star absorbs all matter”).

4. take in – (visit for entertainment; “take in the sights”).

5. collect, take in – (call for and obtain payment of; “we collected over a million dollars in outstanding debts”; “he collected the rent”).

190
6. watch, view, see, catch, take in – (see or watch; “view a show on television”; “This program will be seen all over the world”; “view an exhibition”; “Catch a show on Broadway”; “see a movie”).
7. receive, take in, invite – (express willingness to have in one’s home or environs; “The community warmly received the refugees”).
8. take in, gather in – (fold up; “take in the sails”).
9. absorb, assimilate, ingest, take in – (take up mentally; “he absorbed the knowledge or beliefs of his tribe”).
10. gain, take in, clear, make, earn, realize, realise, pull in, bring in – (earn on some commercial or business transaction; earn as salary or wages; “How much do you make a month in your new job?”; “She earns a lot in her new job”; “this merger brought in lots of money”; “He clears $5,000 each month”).
11. catch, take in, overhear – (hear, usually without the knowledge of the speakers; “We overheard the conversation at the next table”).
12. take in, take up – (accept; “The cloth takes up the liquid”).
13. absorb, suck, imbibe, soak up, sop up, suck up, draw, take in, take up – (take in, also metaphorically; “The sponge absorbs water well”; “She drew strength from the minister’s words”).
14. take in, sop up, suck in, take up – (take up as if with a sponge).
15. consume, ingest, take in, take, have – (serve oneself to, or consume regularly; “Have another bowl of chicken soup!”; “I don’t take sugar in my coffee”).
16. adopt, take in – (take into one’s family; “They adopted two children from Nicaragua”).
17. take in – (make (clothes) smaller; “Please take in this skirt–I’ve lost weight”).

**take off**

1. depart, part, start, start out, set forth, set off, set out, take off – (leave; “The family took off for Florida”).
2. take off – (take away or remove; “Take that weight off me!”).
3. take off, lift off – (depart from the ground; “The plane took off two hours late”).
4. take off, take time off – (take time off from work; stop working temporarily).
5. take off – (mimic or imitate in an amusing or satirical manner; “This song takes off from a famous aria”).
6. take off – (remove clothes; “take off your shirt–it’s very hot in here”).
7. get off the ground, take off – (get started or set in motion, used figuratively; “the project took a long time to get off the ground”).

8. take off – (prove fatal; “The disease took off”).

9. subtract, deduct, take off – (make a subtraction).

take on

1. assume, acquire, adopt, take on, take – (take on a certain form, attribute, or aspect; “His voice took on a sad tone”; “The story took a new turn”; “he adopted an air of superiority”; “She assumed strange manners”; “The gods assume human or animal form in these fables”).

2. assume, adopt, take on, take over – (take on titles, offices, duties, responsibilities; “When will the new President assume office?”).

3. undertake, tackle, take on – (accept as a challenge; “I’ll tackle this difficult task”).

4. accept, admit, take, take on – (admit into a group or community; “accept students for graduate study”; “We’ll have to vote on whether or not to admit a new member”).

5. meet, encounter, play, take on – (contend against an opponent in a sport, game, or battle; “Princeton plays Yale this weekend”; “Charlie likes to play Mary”).

take out

1. take out, move out, remove – (cause to leave; “The teacher took the children out of the classroom”).

2. unpack, take out – (remove from its packing; “unpack the presents”).

3. take away, take out – (take out or remove; “take out the chicken after adding the vegetables”).

4. take out – (obtain by legal or official process; “take out a license”; “take out a patent”).

5. ask out, invite out, take out – (make a date; “Has he asked you out yet?”).

6. take out – (remove something from a container or an enclosed space).

7. take out, buy food – (purchase prepared food to be eaten at home).

8. withdraw, draw, take out, draw off – (remove (a commodity) from (a supply source); “She drew $2,000 from the account”; “The doctors drew medical supplies from the hospital’s emergency bank”).
9. draw, pull, pull out, get out, take out – (bring, take, or pull out of a container or from under a cover; “draw a weapon”; “pull out a gun”; “The mugger pulled a knife on his victim”).

10. draw, take out – (take liquid out of a container or well; “She drew water from the barrel”).

11. extract, pull out, pull, pull up, take out, draw out – (remove, usually with some force or effort; also used in an abstract sense; “pull weeds”; “extract a bad tooth”; “take out a splinter”; “extract information from the telegram”).

12. take out, take away – (buy and consume food from a restaurant or establishment that sells prepared food; “We’ll take out pizza, since I am too tired to cook”).

13. excerpt, extract, take out – (take out of a literary work in order to cite or copy).

14. exclude, except, leave out, leave off, omit, take out – (prevent from being included or considered or accepted; “The bad results were excluded from the report”; “Leave off the top piece”).

**take over**

1. assume, usurp, seize, take over, arrogate – (seize and take control without authority and possibly with force; take as one’s right or possession; “He assumed to himself the right to fill all positions in the town”; “he usurped my rights”; “She seized control of the throne after her husband died”).

2. assume, adopt, take on, take over – (take on titles, offices, duties, responsibilities; “When will the new President assume office?”)

3. take over, relieve – (free someone temporarily from his or her obligations).

4. bear, take over, accept, assume – (take on as one’s own the expenses or debts of another person; “I’ll accept the charges”; “She agreed to bear the responsibility”).

5. take over, buy out, buy up – (take over ownership of; of corporations and companies).

6. repeat, take over – (do over; “They would like to take it over again”).

7. adopt, borrow, take over, take up – (take up and practice as one’s own).

8. absorb, take over – (take up, as of debts or payments; “absorb the costs for something”).

**take up**

1. take up – (pursue or resume; “take up a matter for consideration”).
2. take up, latch on, fasten on, hook on, seize on – (adopt; “take up new ideas”).
3. take up – (turn one’s interest to; “He took up herpetology at the age of fifty”).
4. take up – (take up time or space; “take up the slack”).
5. start, take up – (begin work or acting in a certain capacity, office or job; “Take up a position”; “start a new job”).
6. adopt, borrow, take over, take up – (take up and practice as one’s own).
7. assume, take, strike, take up – (occupy or take on; “He assumes the lotus position”; “She took her seat on the stage”; “We took our seats in the orchestra”; “She took up her position behind the tree”; “strike a pose”).
8. sorb, take up – (take up a liquid or a gas either by adsorption or by absorption).
9. scoop, scoop out, lift out, scoop up, take up – (take out or up with or as if with a scoop; “scoop the sugar out of the container”).
10. take in, take up – (accept; “The cloth takes up the liquid”).
11. absorb, suck, imbibe, soak up, sop up, suck up, draw, take in, take up – (take in, also metaphorically; “The sponge absorbs water well”; “She drew strength from the minister’s words”).
12. take in, sop up, suck in, take up – (take up as if with a sponge).
13. resume, take up – (return to a previous location or condition; “The painting resumed its old condition when we restored it”).

take aback

1. shock, stun, floor, ball over, blow out of the water, take aback – (surprise greatly; knock someone’s socks off; “I was floored when I heard that I was promoted”).

take advantage

1. capitalize, capitalise, take advantage (draw advantages from) “he is capitalizing on her mistake”; “she took advantage of his absence to meet her lover”.
2. trespass, take advantage (make excessive use of) “You are taking advantage of my good will”; “She is trespassing upon my privacy”.

take care
1. take care (be careful, prudent, or watchful) “Take care when you cross the street!”.

2. take care, mind (be in charge of or deal with) “She takes care of all the necessary arrangements”.

3. attend, take care, look, see (take charge of or deal with) “Could you see about lunch?”; “I must attend to this matter”; “She took care of this business”.
Appendix C

FrameNet Annotation Example
for the verb *take*

Examples of annotated sentences for the verb *take*, taken from the FrameNet database, are given below.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<root>
<lexunit-annotation ID="1149" name="take.v" frame="Removing"
    pos="V" incorporatedFE=""/>
<definition>COD: remove from a place</definition>
<lexemes>
<lexeme ID="39414" pos="V" breakBefore="false" headword="false">take</lexeme>
</lexemes>
<annotationSet ID="1152132" status="MANUAL">
<layers>
    <layer ID="6782399" name="FE" rank="1">
        <labels>
            <label name="Agent" ID="21133841" start="0" end="7"/>
            <label name="Theme" ID="21133848" start="14" end="22"/>
            <label name="Source" ID="21133854" start="24" end="40"/>
        </labels>
    </layer>
    <layer ID="6782400" name="GF">
        <labels>
            <label name="Ext" ID="21133843" start="0" end="7"/>
            <label name="Obj" ID="21133850" start="14" end="22"/>
            <label name="Dep" ID="21133856" start="24" end="40"/>
        </labels>
    </layer>
    <layer ID="6782401" name="PT">
        <labels>
            <label name="NP" ID="21133842" start="0" end="7"/>
            <label name="NP" ID="21133849" start="14" end="22"/>
            <label name="PP" ID="21133855" start="24" end="40"/>
        </labels>
    </layer>
    <layer ID="6782402" name="Sent">
        <labels>
            <label name="sense1" ID="29678711" itype="APos"/>
        </labels>
    </layer>
</layers>
```

196
The maid took the plate from her mistress as if it were hot.
Once they had cooled completely we took them out carefully so we did n't disturb the curls too much.

NOTE Some drugs can be dangerous when they are taken at the same time as alcohol.
The doctor at the hospital told me about the pill, but said you have to take them regularly and if you forget even once, then you'll get pregnant.
You took a taxi and I was discovered with a length of picture cord around my neck. 
He took a train back to London, and a taxi over to Charing Cross.
Appendix D

Romanian imported FrameNet Annotation Example

```xml
<?xml version="1.0" encoding="UTF-8"?>
<subcorpus>
  <annotationSet ID="3" status="AUTO">
    <layers>
      <layer name="Target">
        <labels>
          <label name="Target" words="deveni" />
        </labels>
      </layer>
      <layer name="FE" rank="1">
        <labels>
          <label name="Entity" words="Distanțe pentru care este pregătită placerea drumului" />
          <label name="Final_state" words="clar" />
          <label name="Time" words="el și Esau se vor întâlni" />
        </labels>
      </layer>
      <layer name="GF">
        <labels>
          <label name="Ext" words="Distanțe pentru care este pregătită placerea drumului" />
          <label name="Dep" words="clar" />
          <label name="Dep" words="el și Esau se vor întâlni" />
        </labels>
      </layer>
      <layer name="PT">
        <labels>
          <label name="NP" words="Distanțe pentru care este pregătită placerea drumului" />
          <label name="AJP" words="clar" />
          <label name="Sfin" words="el și Esau se vor întâlni" />
        </labels>
      </layer>
      <layer name="Other" />
      <layer name="Sent" />
      <layer name="Verb" />
    </layers>
  </annotationSet>
</subcorpus>
```

Distanțe pentru care este pregătită placerea drumului vor deveni clare când el și Esau se vor întâlni.
Mănăstirea de maici Nunnery Lane s-a implicat în planurile pentru mănăstirea Carmelite de la Mafeking după cele mai cinci ele germane au fost abordate de episcopul de Kimberley al Africii de sud și au luat legătura cu Darlington.
În Marea Britanie a devenit aparent în mod crezător faptul că schimbarea organizațională în școli nu a fost suficientă pentru a garanta schimbare în atitudini sociale stabilite.
Dar am înțeles că comitetul de cricket a votat 4−1 pentru a nominaliza pe batsman-ul Richardson pentru un contract de un an ca să completeze locul vacant până când Craig McDermott va deveni disponibil în 1994.
Fruntea lui devenea din ce în ce mai încrumată pe măsură ce lua decizii ca reacții la circumstanțe, în loc să preia controlul situației.
Urma devenea din ce în ce mai puțin distinct și într-o frumoasă vale abruptă.

Evidența grevelor din anii de imediat după război este greu de aplanat cu notiunea că protestul muncitorilor devenea mai puțin intens.
Din faptul că este, într-o oarecare măsură, mai scump decât energia nucleară, cărbunele a devenit deodată mai ieftin pe piață.
Endilă și Carturarul au devenit buni prieteni.

Tortura deținuților politici a răspândit și a devenit sistematică după asasinarea președintelui Sadat în octombrie 1981.
Ei vă zură luminile satului, împrăștiate și obscure, și ofereașă prinsă pe partea de vest care devenise faimoasă prin zvonuri.
Declaration

I herewith declare that the PhD Thesis entitled **Natural Language Processing Using Semantic Frames** is written entirely by me, without the prohibited assistance of third parties and without making use of aids other than those specified. This paper has not previously been presented in identical or similar form to any other Romanian or foreign examination board. I also declare that notions taken over directly or indirectly from other sources have been identified as such, respecting the intellectual property rights:

- all text fragment exactly reproduced, even if translated from another language, are quoted, having the precise indication of the source;
- the rephrases of texts written by other authors contains the precise indication of the source;
- source code, images, or other information taken from open source projects or other sources are used according to copyright disclaimers and contain the precise references;
- the summarization of other authors ideas contains precise reference to the original text.

Declarație

Prin prezenta declar că teza de doctorat cu titlul **Natural Language Processing Using Semantic Frames** este scrisă de mine și nu a mai fost prezentată niciodată la o altă facultate sau instituție de învățământ superior din țară sau străinătate. De asemenea, declar că toate sursele utilizate, inclusiv cele preluate de pe Internet, sunt indicate în lucrare, cu respectarea regulilor de evitare a plagiatului:

- toate fragmentele de text reproduse exact, chiar și în traducere proprie din altă limbă, sunt scrise între ghilimele și dețin referința precisă a sursei;
- reformularea în cuvinte proprii a textelor scrise de către alții autori deține referința precisă;
- codul sursă, imagini etc. preluate din proiecte open-source sau alte surse sunt utilizate cu respectarea drepturilor de autor și dețin referințe precise;
- rezumarea ideilor altor autori precizează referința precisă la textul original.