Computational lexicography, morphology and syntax

Diana Trandabăț

Academic year: 2022-2023

About words…

- Words in natural languages usually encode many pieces of information:
  - What the word “means” in the real world
  - What categories, if any, the word belongs to
  - What is the function of the word in the sentence
- Nouns: How many?, Do we already know what they are?, How does it relate to the verb?, …
- Verbs: When, how, who,…

Why do we care about words?

- Many language processing applications need to extract the information encoded in the words.
- Parsers which analyze sentence structure need to know/check agreement between
  - subjects and verbs
  - Adjectives and nouns
  - Determiners and nouns, etc.
- Information retrieval systems benefit from knowing what the *stem* of a word is
- Machine translation systems need to analyze words to their components and generate words with specific features in the target language (e.g. *compound words*)

Morphology - definition

- **Morphology** is concerned with the ways in which words are formed from basic sequences of phonemes.
- The study of the internal structure of words

History

- Well-structured lists of morphological forms of Sumerian words were attested on clay tablets from Ancient Mesopotamia and date from around 1600 BC; e.g. (Jacobsen 1974: 53-4):
  - badu ‘he goes away’
  - baddun ‘I go away’
  - bašidu ‘he goes away to him’
  - bašiduun ‘I go away to him’

Morphology - types

- Two types are distinguished:
  - inflectional morphology
  - derivational morphology
- Words in many languages differ in form according to different functions:
  - nouns in singular and plural *(table* and *tables)*
  - verbs in present and past tenses *(likes* and *liked),* etc.

Inflectional morphology

- **Inflectional morphology** - the system defining the possible variations on a root (or base) form, which in traditional grammars were given as ‘paradigms’
  - Ex. Latin *dominus, dominum, domini, domino,* etc.
  - The root *domin-* is combined with various endings *(-us, -um, -i, -o,* etc.), which may also occur with other forms: *equus, servus,* etc.
  - English is relatively poor in inflectional variation:
    - most verbs have only *-s, -ed* and *–ing* available;
  - Romanian language is much richer.

Inflectional morphology

- Languages - according to the extent to which they use inflectional morphology:
  - so-called **isolating** languages (Chinese), which have almost no inflectional morphology;
  - **agglutinative** languages (Turkish), where inflectional suffixes can be added one after the other to a root,
  - **inflecting** languages (Latin), - simple affixes convey complex meanings: for example, the *-o* ending in Latin *amo* (‘I love’) indicates person (1st), number (singular), tense (present), voice (active) and mood (indicative).
  - **polysynthetic** languages (Eskimo) is said to be an example, where most of the grammatical meaning of a sentence is expressed by inflections on verbs and nouns.

Isolating languages

- Isolating languages do not (usually) have any bound morphemes
  - Mandarin Chinese
  - *Gou bu ai chi qingcai* (dog not like eat vegetable)
  - This can mean one of the following (depending on the context)
    - *The dog doesn’t like to eat vegetables*
    - *The dog didn’t like to eat vegetables*
    - *The dogs don’t like to eat vegetables*
    - *The dogs didn’t like to eat vegetables.*
    - *Dogs don’t like to eat vegetables.*

Agglutinative Languages

- (Usually multiple) Bound morphemes are attached to one (or more) free morphemes, like beads on a string.
  - Turkish/Turkic, Finnish, Hungarian
  - Swahili, Aymara
- Each morpheme (usually) encodes one "piece" of linguistic information.

Polysynthetic Languages

- Use morphology to combine syntactically related components (e.g. verbs and their arguments) of a sentence together
  - Certain Eskimo languages, e.g., Inuktikut
  - *qaya:liyu:lumi:* he was excellent at making kayaks

Derivational morphology

- **Derivational morphology**: formation of root (inflectable) forms from other roots, often of different grammatical categories (see below).
  - *nation* (noun) -> *national* (adjective) -> *nationalise* (verb)
  - *nation* (noun) -> *national* (adjective) -> *nationalism* (noun)
  - *nation* (noun) -> *national* (adjective) -> *nationalist* (noun).
  - *nation* (noun) -> *national* (adjective) -> *denationalisation* (noun)

Word-form

- **Word form**: A concrete word as it occurs in real speech or text.
- For our purposes, word is a string of characters separated by spaces in writing.

- **Lemma:** A distinguished word form from a set of morphologically related forms, chosen by convention (e.g., nominative singular for nouns, infinitive for verbs) to represent that set. Also called the canonical/base/dictionary/citation form. For every form, there is a corresponding lemma.

Lexeme

- **Lexeme**: An abstract entity, a dictionary word; it can be thought of as a set of word-forms. Every form belongs to one lexeme, referred to by its lemma.
- For example, in English, *steal, stole, steals, stealing* are forms of the same lexeme steal; steal is traditionally used as the lemma denoting this lexeme.
- **Paradigm**: The set of word-forms that belong to a single lexeme.

Paradigm

- The paradigm of the Romanian *insulă*

	singular	plural
nominative	insulă	insule
accusative	insulă	insule
genitive	insulei	insulelor
dative	insulei	insulelor
vocativ	insulă	insule

Computational morphology

- Computational morphology deals with
  - developing theories and techniques for
  - computational analysis and synthesis of word forms.
- Analysis: Separate and identify the constituent morphemes and mark the information they encode
- Synthesis (Generation): Given a set constituent morphemes or information be encoded, produce the corresponding word(s)

Computational Morphology -Analysis

- Computational morphology deals with
  - developing theories and techniques for
  - computational analysis and synthesis of word forms.
- Extract any information encoded in a word and bring it out so that later layers of processing can make use of it

stopping	⇒ stop+Verb+Cont
happiest	⇒ happy+Adj+Superlative
went	⇒ go+Verb+Past
books	⇒ book+Noun+Plural
	⇒ book+Verb+Pres+3SG.

Computational Morphology -Generation

- In a machine translation applications, one may have to generate the word corresponding to a set of features
  - stop+Past ⇒ stopped
  - cânta+Past+1Pl ⇒ cântaserăm/cântasem

    +2Pl ⇒ cântaserăți/cântasei

Computational Morphology-Analysis

- Input raw text
- Segment / Tokenize

Pre-processing

- Analyze individual words
- Analyze multi-word constructs
- Disambiguate Morphology

Morphological processing

- Syntactically analyze sentences

Syntactic processing

Examples of applications

- Spelling Checking
  - Check if words in a text are all valid words
- Spelling Correction
  - Find the correct words “close” to a misspelled word.
- For both these applications, one needs to know what constitutes a valid word in a language.
  - Rather straightforward for English

Examples of applications

- Grammar Checking
  - Checks if a (local) sequence of words violates some basic constraints of language (e.g., agreement)
- Text-to-speech
  - Proper stress/prosody may depend on proper identification of morphemes
- Machine Translation (especially between closely related languages)

Morphological Ambiguity

- Morphological structure/interpretation is usually ambiguous
  - Part-of-speech ambiguity
    - book (verb), book (noun)
  - Morpheme ambiguity
    - +s (plural) +s (present tense, 3rd singular)
- Segmentation ambiguity
  - Word can be legitimately divided into morphemes in a number of ways

Morphological Ambiguity

- The same surface form is interpreted in many possible ways in different syntactic contexts. In French, *danse* has the following interpretations:
- danse+Verb+Subj+3sg (lest s/he dance)
- danse+Verb+Subj+1sg (lest I dance)
- danse+Verb+Imp+2sg ((you) dance!)
- danse+Verb+Ind+3sg ((s/he) dances)
- danse+Verb+Ind+1sg ((I) dance)
- danse+Noun+Fem+Sg (dance)

Morphological Disambiguation

- Morphological Disambiguation or Tagging is the process of choosing the "proper" morphological interpretation of a token in a given context.

He can can the can.

Morphological Disambiguation

- He can can the can.

- Modal
- Infinitive form
- Singular Noun

- Non-third person present tense verb
  - We can tomatoes every summer.

Morphological disambiguation

- These days standard statistical approaches (e.g., Hidden Markov Models) can solve this problem with quite high accuracy.
- The accuracy for languages with complex morphology/ large number of tags is lower

Implementation Approaches for Computational Morphology

- List all word-forms as a database
- Heuristic/Rule-based affix-stripping
- Finite State Approaches

Why is the Finite State Approach Interesting?

- Finite state systems are mathematically well-understood, elegant, flexible.
- Finite state systems are computationally efficient.
- For typical natural language processing tasks, finite state systems provide compact representations.
- Finite state systems are inherently bidirectional

Romanian morphology

- specific characteristics that contribute to the richness of the language, but are also a challenge for NLP.
- Romanian's inflection is quite rich.
- For nouns, pronouns and adjectives – 5 cases and 2 numbers.
- Pronouns can have stressed and unstressed forms
- Nouns and adjectives can be defined or undefined.
- Verbs – 2 numbers, each with 3 persons and 5 synthetic tenses, plus infinitive, gerund and participle forms.
- Average: noun - 5 forms, personal pronoun - 6 forms, adjective - 6 forms, verb > 30 forms.
- Besides morphologic affixes, phonetic alternations inside the root are also possible with inflected words.

Grammar reminder - nouns

- 5 cases and 2 numbers
- Nouns can be defined or undefined

- Choose a noun and derivate it!

- Bonus for finding one with phonetic alternations inside the root ☺

Grammar reminder - adjectives

- 5 cases and 2 numbers
- Adjectives can be defined or undefined

- Choose an adjective and derivate it!

- Bonus for finding one with phonetic alternations inside the root ☺

Grammar reminder - pronouns

- 5 cases and 2 numbers
- Pronouns can have stressed and unstressed form

- Choose a pronoun and derivate it!

Grammar reminder - verbs

- Verbs – 2 numbers, each with 3 persons and 5 synthetic tenses, plus infinitive, gerund and participle forms.

- Choose a verb and derivate it!

- Bonus for finding one with phonetic alternations inside the root ☺

How to read „morphology”

- Știe.
- Knows-he/she/it
- ‘He/She/It knows. ’

- I$_i$ l$_j$ –am dat mamei$_i$ pe Ion la telefon.
- Dat. cl. Acc. masc. cl. have-I given to-mother John over the phone.
- ‘I gave John to my mother on the phone.’

Now its your tour!

- Write in the same form the translation for the sentence:

Ion le-a multumit prietenilor pentru cadou.

Until next week…

“My definition of dictionary can’t be found in the dictionary.

Dictionary - A linguistic prison, confining words to well-defined cells, with little chance of parole.”

Jarod Kintz -

How to construct a coffin with six karate chops