Linking Book Characters

Toward A Corpus Encoding Relations Between Entities

Dan Cristea
"Alexandru Ioan Cuza" University of Iași, Department of Computer Science (UAIC-FII)
Institute for Computer Science, Romanian Academy, Iași branch
dcristea@info.uaic.ro

Eugen Ignat
UAIC-FII
eugen.ignat@info.uaic.ro

***Abstract*— What does a novel bring to a reader? What can it bring to a machine? Are there chances that a machine will decipher the messages a book expresses in free language? Part of the content of a text is encoded in relations between entities. In order to decode them, algorithms make use of learning techniques in which the training is guided by corpora that make explicit entities and relations. The creation of a gold corpus to be used in training and evaluation is therefore of a primary concern. This paper proposes annotation conventions and methodological prerequisites for the creation of a corpus that puts in evidence characters in a book and relations that are mentioned as holding between them, of the types: anaphoric, affective, kinship and social. The language under investigation is Romanian and the type of text used is fiction, but the proposed conventions are thought to be applicable to any language and type of text.**

Keywords*— *entity linking; anaphoric relations; semantic relations; annotated corpora; annotation conventions; content analysis; text analytics; text understanding; XML.

I. INTRODUCTION

Books are written for humans, not for machines. Same as music, as paintings, as theatre or as dance, books are intended to impress us, but more than all these other forms of art, they can also bring us knowledge. We are able to learn out of books about past events, behaviours of famous people, or places we've never been in yet. Non trivially, books influence our personality, change our perspective over life, modify our way of thinking. To what extend can we exploit the books' content more than by keeping them under the eyes and reading them? To what extend could a machine touch the content hidden in books in order to present us in an organised manner? In what way would work a technology able to conceptualise parts of the knowledge encrypted in books? This paper is a preliminary study toward these goals.

In general, the entity linking task is defined as the process of linking named entities found in unstructured texts to records of a knowledge base. The task is thought to be important for several information extraction and natural language processing applications, such as search in unstructured texts and summarization.

Parts of this paper are based on a project proposal that presents a technology, called *MappingBooks*, intended to connect the text of a book, as well as its readers, with information extracted from the Web and in the reality around. A *MappedBook* is a book connected with locations/events in the virtual and real world and sensible to the instantaneous location (as seized by the mobile/tablet) of a reader. The information made available could possibly be different depending on the moment and the place of the reader. For instance, if the user reads in a touristic guide about a museum and her/his momentary position happens to be in its proximity, the system will try to find out from the web the open hours of the museum. Then, in case there is still time for a visit, it will signal to the user about this possibility. This way, the text of the touristic guide comes to life and becomes intimately connected to the reader. Features of interactive social gaming could put in correspondence children emerged into a learning-by-visiting activity related to Geography or tourists' journeys.

The *MappingBooks* technology puts in the same melting pot known and new techniques to create multi-dimensional mash-ups combining textual, geographical and even temporal information in order to present them adequately to a reader, intermediated by a specially-designed human-computer interface. The links should be sensible to the context of the mentions in the book, the moment the user initiates an access and the current location of the user. The technology is intended to make heavy use of entity linking techniques, making possible to spot the mentioned entities in the book (persons and locations) in the real and the virtual world.

Out of the *MappingBooks* proposal (that involves many other issues related to mixed reality and geo-informatics) we refer here only to the text analytics and content extraction machinery. We take the *MappingBooks* setting to exemplify possible applications of an entity linking technology.

In this paper we discuss a representational mechanism that makes possible the text analytics and entity linking techniques that should be put at the core of a technology as the one described in *MappingBooks*. Especially, we concentrate on building a corpus annotated with information about a diversity of types of entity mentions and four types of relations (anaphoric, affective, kinship and social), out of which a system would learn to detect by itself the searched for entity mentions and the relations inter-connecting them.

These are a number of features that characterise our approach: 1). the mentions we want to connect are not necessarily entities with names, but general noun phrases (with pronominal, common or proper nouns heads and modifiers); 2). there are no preliminary records about the entities linked, so the knowledge base evolves from scratch; 3). we try to connect in anaphoric coreferential chains all mentions of the same entity, each such chain being attached to exactly one entity in the universe of the text; 4). we are looking only for a well established set of relations these entities are involved in; 5). the texts we investigate are fiction books, therefore, on one hand, for each analysis, the reference area is well delimited and, on the other hand, the form of expression is totally unrestricted.

The paper includes the following sections: Section II gives a brief state of the art of the emerging entity linking domain, mentioning some of its most known methods and basic techniques. Section III presents a proposal for annotating a corpus of entities and relations and Section IV includes a discussion and some concluding remarks.

II. PRIOR ART AND BASIC LEVEL PROCESSING

In entity linking, the extracted data are stored in a knowledge base (KB), which is continuously evolving during the process. New extractions must be merged with already existent information, which imposes care for determining when a match occurs or when new entries should be created in the KB. Depending on the application, the source data is a closed collection (for instance a book) or is open – the whole web, or only parts of it (Wikipedia, DBpedia, touristic pages, mass-media reflected on the Web, social-media, blogs, etc.).

The most important challenges in entity linking address: name variations (different text strings in the source text refer the same KB entity), ambiguities (there is more than one entity in the KB a string can refer to) and absence (there is no entity description in the KB to which a string in the source text representing an entity could possibly match). Bunescu and Pasca [1] and Cucerzan [2] presented important pioneering work in this area. Cucerzan does not handle absences while Bunescu and Pasca address them by learning a threshold. Some approaches use supervised machine learning. For instance, Rao et al. [3] score entities contained in the KB for a possible match to the query entity.

NELL (Never-Ending Language Learning) is a project[1] intending to make a computer learn to "read the web" [4, 5]. It browses hundreds of millions of web pages and applies several extraction methods to detect plausible facts involving entities and categories. These facts are then used to enhance its technique, such that, in time, it becomes more and more competent. The initial input of NELL has been an ontology including several hundreds of categories and binary relations. The "beliefs" NELL is continuously extracting from the web are used as a self-supervised collection of training examples. Over its exploitation, the extracted knowledge had to be manually revised several times as to eliminate erroneous facts, which, let to proliferate, would have deteriorate its future behaviour.

The knowledge NELL extracts is of a static nature, since no temporal relations are taken into consideration. Other approaches address the challenge to determine time constraints as general sequencing knowledge, out of which to infer the moments or the periods of time some particular events had occurred [6].

The 2012 Text Analysis Conference (TAC) launched the Knowledge Base Population (KBP) Cold Start task, requiring systems to take a set of documents and produce a comprehensive set of <Subject, Predicate, Object> triples that encode relationships involving named-entities. TAC-2012 used a fixed schema of 42 relations and their logical inverses. One of systems participating in this task [7] report between 30-80% accuracy, by manually examining ten random samples for each relations in the absence of a gold corpus. The system, called KELVIN, integrates a pipeline of processing tools, among which a basic tool is the BBN's SERIF (Statistical Entity & Relation Information Finding). SERIF [8] does named-entities identification and classification by type and subtype, intra-document co-reference analysis, including named, nominal and pronominal mentions, sentence parsing, in order to extract intra-sentential relations between entities, and detection of certain types of events. SERIF provides a considerable suite of document annotations that have been used as basis for building an initial KB.

But an accurate entity linking technique is dependent on a diversity of NLP mechanisms, which should work well in correlation. An aphorism which circulates in NLP circles, referring only to a part of the domain, anaphora resolution, says that once you have done everything in NLP you solved anaphora for free. Ad-hoc linking techniques are on the class of one-document and cross-document anaphora resolution [9, 10, 11, 12]. RARE, the UAIC-FII's system of anaphora resolution, relying on a mixed approach which combines symbolic rules with learning techniques, has given good results for Bulgarian, German, Greek, English, Polish, and Romanian [13, 14, 15, 16].

Research on detection of content features at the pre-syntactic, syntactic and semantic levels, such as noun phrases, dependency relations and semantic roles, were the subject of many studies. Here we mention only a few, pursued within the NLP-Group@UAIC-FII, reporting technologies that correlate well with the issue discussed in this paper: noun phrase chunking [17], dependency parsing [18], clause splitting [19, 16][2]. Textual entailment is a problem of central concern in mastering the diversity of expressions with similar content [20, 21] and techniques of semantic roles labelling [22] are a source of good hints is searching verbal relations between entities. Also, an application of entity linking, the issue of extracting patterns of geographic knowledge in order to map a textual description of a travel onto GoogleMaps, was studied in two graduation thesis [23, 24].

As in many applications of NLP, entity linking too cannot be done without proper collections of annotated texts, where human linguistic expertise is used to make explicit the deep semantic knowledge of the kind the future technology is

[1] http://rtw.ml.cmu.edu/rtw/

[2] Part of them are active web services at http://nlptools.info.uaic.ro/

supposed to discover by itself. But the success of the attempt to acquire this kind of superior corpus relies very much on the maturity of basic NLP technologies (including at least: tokenization, part-of-speech tagging and lemmatization, noun-phrase chunking, name entity recognition, and segmentation at sentence level), since basic elements are necessary for automatic processing and a massive manual annotation of them is not feasible, because of extremely high costs and time constraints. As such, there are two ways to go further in this enterprise: either a manually annotated corpus including basic elements already exists and it is further annotated by experts at higher levels, or a virgin text is chosen, on which basic linguistic processing is launched, and human expertise involving the superior levels is added only on top of the automatically marked elements. The second methodology has the advantage that the gold corpus and the test corpus (as well as the future texts processed by the technology) are characterised by similar annotation accuracies at basic levels (given by the basic NLP chains). If this decision for building the gold corpus is adopted, extracting content from free texts is usually the ultimate step of a long process addressing basic text analytics.

The techniques doing this processing seem to have reached technological maturity, as proved by recent European projects CLARIN[3], METANET4U[4], ATLAS[5] (to mention only some in which the authors have been involved). For instance, Anechitei et al. [16] report precision of base components of the ATLAS NLP chain[6] in the range 80% – 97% for Romanian and English. Moreover, the same research shows how individual tools can be easily combined in processing chains by using UIMA-based interfaces [25]. Similar type of processing is used for the aggregation of language processing pipelines into NLP applications in METANET4U, where the U-Compare CAS interface [26], a dialect of UIMA, has been used.

III. ANNOTATION CONVENTIONS

As mentioned in Section I, in the research described in this paper we are concerned in proposing adequate annotation conventions that would support the development of a gold corpus. We discuss in this section two levels of annotation above the basic markings which includes token and noun phrase boundaries and word level morphological information. One level puts in evidence entities; the other marks relations between them.

A. Annotating entities

Syntactically, at the text level, the entities are signalled by noun phrases. Their annotation is important because they can participate in relations. To refer to them we will use a term which is common in anaphora studies: referential expressions (REs). Here are some rules for recognising REs:

- REs have nominal or pronominal heads, may include modifiers (determiners, adjectives, numerals, genitival constructions, prepositional phrases), but they do not extend also over relative clauses.

- We will consider that **each RE evokes an entity**, which is part either of the same text as the RE itself, or of the virtual world (as a reflection of the real world). By borrowing a term from Centering [20], we will say that each RE *realises* an entity.

- If two nested noun phrases share the same syntactic head, only the largest is retained as a RE. If two REs are intersectable, they are necessarily nested. Nested REs should have distinctive heads, which also means that they cannot realise identical entities (or, following the terminology from section III.B, they cannot co-refer). For instance *the lady with the red hat* and *the red hat* realise distinct entities.

- In general, entities have types: PERSON, LOCATION, ORGANISATION, DATE, ADDRESS, OTHER, but in the exercise described in this paper, we are interested only in PERSON type entities (including groups).

- When entities have associated descriptions, it is important to distinguish between identification and characterisation descriptions. Only identification descriptions are included in REs. For instance, in *acest bărbat, stricat până-n măduva oaselor*[7] *<that man, corrupted to the marrow of his bones>* the span *stricat până-n măduva oaselor <corrupted to the marrow of his bones>* is a characterisation description. It does not help in the identification of the entity, of a man among many. On the other hand, in *the man with straw hat* the span *straw hat* is an identification description. It contributes in distinguishing one man in a group and it must be part of the RE.

- We say that an entity is *included* if it is not realised in the text. In Romanian, included entities appear only in the position of subject. We annotate such entities only if they participate in a relation that is among those decided to be annotated. For instance, in *dar îl și iubeau din tot sufletul <but loved him with the whole soul>*, the subject of *iubeau <loved>* is included. Here, the morpho-syntactic properties of the included entity are preserved in the person and number of the verb. We will annotate 1:[*îl*] and 2:[*iubeau*; REALISATION=INCLUDED]) *din tot sufletul*, therefore, the verb as the second entity, participating in an affective relationship (`love`, as described in Section III.B).

We adopted an XML notation for entities, that shows the span over which it extends and the type. As noted, an entity marking should extend over a noun phrase, which includes the component tokens (words). The XML element for a word is `W` and, apart from an `ID`, it includes morpho-syntactic attributes, among which – `LEMMA`. The element denoting an entity is `ENTITY`, with attributes: `ID` and `TYPE` (and optionally, `HEAD`). As explained, for included subjects the verb is

[3] http://www.clarin.eu/external/
[4] http://metanet4u.eu/
[5] http://www.atlasproject.eu/
[6] http://www.atlasproject.eu/atlas/file/911996c9-0b0a-48ef-868a-7a483081e8f0/13ccaaa0-a0c0-11e0-8264- 0800200c9a66/Finalreport.txt

[7] Most of the examples in this paper are taken from the Romanian version of Quo Vadis, by Henryk Sienkiewicz, București, 1967, in the translation from Polish by Remus Luca and Elena Lință. The English equivalents have been searched in the translation from Polish by Jeremiah Curtin (project GUTENBERG eBook Quo Vadis). All examples are marked in *italics* and the English equivalents follow the Romanian examples in angular brackets, like this: *Romanian example <English equivalent>*.

annotated instead, which adds to the `ENTITY` element also the attribute-value pair: `REALISATION="INCLUDED"`.

B. Annotating anaphoric relations

Anaphora is known to be a relation between two referential expressions: the reference of the *anaphor* depends upon another referential element, usually called *antecedent*. If the element the anaphor depends on precedes the anaphor we call the relation anaphora, otherwise – cataphora. We use the term anaphoric in its large sense, which includes not only strict coreferences (identity of reference), but also other types (part-of, class-of, etc.) – the complete list is given below in this section.

Following are a number of hints that guided the process of annotation of anaphoric relations:

- All anaphoric relations linking non-imbricated entities (imbrication here means overlapping of REs) are directed in the text right-to-left, therefore from the anaphor towards the antecedent. This direction corresponds to the usual interpretation done by humans during reading: the current referential expression is interpreted related to entities already mentioned in the unfolded text, therefore to the left of the RE under scrutiny. The same direction is considered also in case of cataphora, see [13].
- The anaphoric relations between imbricated entities, by convention, are annotated from the most largest RE towards the included one, for instance: 1:[*femei din* 2:[*societatea înaltă*]] <1:[*women of* 2:[*the high society*]]> => [1] `member-of` [2].

The following types of anaphoric relations are annotated:

- `coref` (coreferentiality, symmetrical). Example: 1:[*Acteea*]... 2:[*tânara libertă*] <1:[*Acte*]… 2:[*the young freedwoman*]> => [2] `coref` [1];
- `member-of` (from one element to the group the element is a member of), see the previous example with *femei din societatea înaltă*;
- `has-as-member` (the inverse of `member-of`: from a group to one element belonging to it), for instance: "1:[*Petronius*]... 2:[*Vinicius*]... 3:[*amândurora*]" <1:[*Petronius*]... 2:[*Vinicius*]... *to* 3:[*both of them*]> => [3] `has-as-member` [1]; [3] `has-as-member` [2]; 1:[*X*], 2:[*Y*], 3:[*Z*] *și* 4:[*alte* 5:[*femei din* 6:[*societatea înaltă*]]] <1:[*X*], 2:[*Y*], 3:[*Z*] *and* 4:[*other* 5:[*woman of* 6:[*the high society*]]]> => [5] `has-as-member` [1], etc.; 1:[*Ursus*]… 2:[*Ligia*]… 3:[*voi*] <1:[*Ursus*]… 2:[*Ligia*]… 3:[*you*$_{PL}$]> => [3] `has-as-member` [1]; [3] `has-as-member` [2];
- `isa` (from an instance to its conceptual class), for example: 1:[*Profesor de matematică*] *s-a* 2:[*făcut*; REALISATION=INCLUDED]*, așa cum a visat încă de copil.* <1:[*A profesor of Maths*] 2:[*he*] *became, as he dreamed since childhood.*[8]> => [2] `isa` [1];
- `class-of` (the inverse of `isa`, therefore from a concept to an instance): 1:[*Ion*] *a devenit* 2:[*profesor de matematică*]. <1:[*John*] *became* 2:[*a teacher of Maths*].> => [2] `class-of` [1];
- `part-of` (a component part of an assembly), for instance: *luându*-1:[*i*] 2:[*mâinile*] <*taking* 2:[1:[*his*] *hands*]> => [2] `part-of` [1][9].
- `has-as-part` (the inverse of `part-of`: X `has-as-part` Y if X has Y as a component part), for instance: 1:[*Ochii*] 2:[*îi*] 3:[*ai*; REALISATION=INCLUDED] *de la tata sau de la mama?* <1:[*The eyes*], 3:[*you*] *have* 2:[*them*] *from your father or your mother?*> => [3] `has-as-part` [1] (and [2] `coref` [1]). Note that there are two ways to read *omul cu mâna în ghips* <*the man with the hand in plaster*>, namely as 1:[*omul cu* 2:[*mâna în ghips*]] <1:[*the man with* 2:[*the hand in plaster*]]> and as 3:[*omul cu* 4:[*mâna*] *în ghips*]] <3:[*the man with* 4:[*the hand*] *in plaster*]> and only in the second interpretation a relation [3] `has-as-part` [4] holds, because only a hand can be part of a man, not also a hand in plaster.
- `subgroup-of` (from a subgroup to a larger group, which includes it), for instance: *Hristos* 1:[*i*]*-a iertat și pe* 2:[*evreii*] *care l-au dus la moarte și pe* 3:[*soldații romani*] *care l-au țintuit pe cruce.* <*Christ has forgiven* (1:[*them*$_{DOUBLING\_CLITIC}$]) 2:[*the Jews*] *who brought him to death, just as* 3:[*the Roman soldiers*] *who have nailed him against the cross.*> => [2] `subgroup-of` [1]; [3] `subgroup-of` [1].
- `has-as-subgroup` (the inverse of `subgroup-of`, i.e. from a group to a subgroup of it), for instance: *Cu ocazia zilei tatălui,* 1:[*băieții*] *s-au reîntâlnit cu* 2:[*familia*]. <*On their father's birthday,* 1:[*the boys*] *came back in* 2:[*the family*].> => [2] `has-as-subgroup` [1].
- `has-name` (the relationship between an entity and its name), for instance: *Atunci* 1:[*"sagatio"*]*, cum numeau* 2:[*aruncarea în sus pe pelerina soldățească*]... <*Then the* 1:[*"sagatio"*]*, as they termed* 2:[*the tossing*]> => [2] `has-name` [1];
- `name-of` (the inverse of `has-name`, i.e. from a name to the entity that cares this name), for instance: *Nero numea* 1:[*aceste incursiuni*] 2:[*pescuit de perle*]. <*Nero termed* 1:[*these ingressions*] 2:[*pearls fishing*].> => [2] `name-of` [1]; 1:[*numele lui* 2:[*Aulus*]] <1:[*name of* 2:[*Aulus*]]> => [1] `name-of` [2].

All types of anaphoric relations mentioned above are doubled by a similar number of cases in which the anaphoricity

[8] The translation in English adopts an unnatural word order, but is left as such to preserve the `isa` relation, same as in Romanian.

[9] Note that the two REs are non-imbricated in the original Romanian and imbricated in the English version. However, in both variants the sense of the relation is conformant with the conventions stated above, in this section.

is dependent on the interpretation of a character, or is doubtful, or not yet realised. To put in evidence such person-mediated interpretations, the name of the relations are complemented with the ending "`-interpret`". We discuss below only some cases:

- `coref-interpret` (coreference by virtue of the interpretation given by a character, following the vision of someone, also a symmetrical relation). Sometimes the `coref-interpret` relation holds between a predicative name and a subject: *Nu avea nici cea mai mică îndoială că* 1:[*lucrătorul acela*] *e* 2:[*Ursus*]. *<He had no doubt that* 1:[*that worker*] *is* 2:[*Ursus*].> => [2] coref-interpret [1]; *L-am prezentat pe* 1:[*acest Glaucus*] *ca pe* 2:[*fiul Iudei*] *și* 3:[*trădător al tuturor creștinilor*]. *<I described* 1:[*Glaucus*] *as* 2:[*a real son of Judas*]*, and* 3:[*a traitor to all Christians*].> => [2] `coref-interpret` [1]; [3] coref-interpret [1]);
- `class-of-interpret` (a relation from a concept to an instance of it, in the vision of someone). Examples: *dar nu ești* 1:[*tu*] 2:[*un zeu*]*?* (*but aren't* 1:[*you*] 2:[*a God*]*?*) => [2] `class-of-interpret` [1]; *dați-mi-* 1:[*o*] *de* 2:[*nevastă*] (*give* 1:[*her*] *to me as* 2:[*wife*]) => [2] `class-of-interpret` [1].

The XML element for marking anaphoric relations is `REFERENTIAL`, with attributes: `ID`, `FROM` (indicating the ID of the anaphor, an `ENTITY` element) `TO` (ID of the antecedent, also an `ENTITY` element) and `TYPE` (one of the types discussed above). The `REFERENTIAL` elements do not mark boundaries of text.

C. Annotating non-anaphoric relations

Before inventorying the types of non-anaphoric relations, we state below a number of principles we adhered to while marking them:

- The marking should indicate the minimal span that includes the two *poles* (arguments of the relation) and the *trigger* (a word making the relation explicit in the text). Usually the name of the relation should be identical with the trigger's lemma, or should be a synonym or a hypernym of it. Among the two poles, one is the *source* (or *actant*) and the other – the *destinator* (or *passive*), such that the relation should be read in the text between the source pole and the destination pole. This criterion indicates the *sense* of the relation and is supposed to disambiguate between a relation and its inverse. It equally applies when the two poles are realised by intersecting or non-intersecting spans of text. Example of non-intersecting poles: 1:[*Ion*] 2:[*o*] 3:[*iubește*] *pe* 4:[*Maria*]. <1:[*John*] 3:[*loves*] 4:[*Mary*].> => [1] `love` [4], trigger: [3] (also [4] `coref` [2]), and of intersecting poles: 1:[2:[*comandantul*] *lui* 3:[*Ion*]] <1:[3:[*John*]'s 2:[*commander*]]> => [1] `superior-of` [3], trigger: [2].
- When one pole is missing, as for instance in case of an included subject, it is replaced in the annotation by the token whose morphological attributes supplements those of the missing element (for instance, the main verb in case of a null subject). Example: *dar* 1:[*îl*] *și* 2:[*iubeau*; REALISATION=INCLUDED]) *din tot sufletul <but* 2:[*loved*;REALISATION=INCLUDED] 1:[*him*] *with the whole soul*> => [2] `love` [1], trigger: [2].

The following types of non-anaphoric relations are annotated:

- Social relations: `superior-of`, `inferior-of` (inverse of `superior-of`), `colleague-with`. Examples: in 1:[*împăratul*] *și* 2:[3:[*curtenii*] 4:[*lui*] *de frunte*] *dormeau încă* <1:[*Cesar*] *and* 2:[4:[*his*] *principal* 3:[*courtiers*]] *were sleeping yet*>, the annotated span marking the social relation `inferior-of` is *curtenii lui de frunte <his principal courtiers*>, because it includes all relevant elements, the two poles and the trigger: [2] `inferior-of` [4], trigger: [3]. Let's note also that [4] is in a `coref` relation with [1], placed outside the span of the social relation.
- Affective relations: `friend-of`, `enemy-of`, `love`, `hate` and `worship`. Examples: 1:[2:[*tovarășii*] 3:[*săi*]] <1:[3:[*his*] 2:[*friends*]]> => [1] friend-of [3], trigger: [2]; 1:[*oamenii aceia*] *nu numai că-și* 2:[*slăveau*] 3:[*zeul*] <1:[*those people*] *not merely* 2:[*honored*] 3:[*their God*]> => [1] `worship` [3], trigger: [2]; 1:[*Ligia*] *îngenunche ca să se* 2:[*roage*] 3:[*altcuiva*]. *<But* 1:[*Lygia*] *dropped on her knees to* 2:[*implore*] 3:[*someone else*].> => [1] `worship` [3], trigger [2].
- Kinship relations: `parent-of` (includes any relation between parents and children), `child-of` (inverse of `parent-of`), `grandparent-of`, `grandchild-of` (inverse of `grandparent-of`), `sibling` (symmetrical, between brothers and sisters), `ant-uncle-of`, `nephew-of` (inverse of `ant-uncle-of`), `cousin-of` (symmetrical, between cousins), `spouse-of` (symmetrical, between husbands), `unknown` (when the kinship type is not mentioned). We do not put in evidence the gender of the two poles, therefore no distinction is made between father, mother, son, daughter, etc. as well as the number of actants and passives. Examples: 1:[2:[*tatăl*] 3:[*lor*]] <1:[3:[*their*] 2:[*father*]]> => [1] parent-of [3], trigger: [2]; 1:[*o* 2:[*rudă*] *de-a lui* 3:[*Petronius*]] *<a relative of Petronius*> => [1] `unknown` [3], trigger: [2].

The XML notations for non-anaphoric relations include the elements: `SOCIAL`, `AFFECTIVE` and `KINSHIP`, all with the attributes: `ID`, `FROM`, `TO`, `TRIGGER`, `TYPE`. The markings should delimitate the boundaries.

Let's discuss below the following span: 1:[Marcus Vinicius] era 2:[3:[fiul] 4:[5:[surorii] 6:[sale] mai mari]], 7:[care], cu ani în urmă, se 8:[căsătorise] cu 9:[10:[tatăl] 11:[acestuia]], 12:[13:[consul] pe vremea lui 14:[Tiberiu]]. <1:[Vinicius] was 2:[the 3:[son] of 4:[6:[his] oldest 5:[sister]]], 7:[who] years before had 8:[married] 9:[11:[his] 10:[father]], 12:[a 13:[man] of consular dignity from the time of

14:[Tiberius]].>. The entities are shown in square brackets and the following relations have to be marked:

- referential: [2] `coref` [1]; [7] `coref` [4]; [10] `coref` [1]; [12] `class-of` [9];

- kinship: [4] `sibling` [6], trigger: [5]; [2] `child-of` [4], trigger: [3]; [7] `spouse-of` [9], trigger:[8]; [9] `parent-of` [11], trigger:[10];

- social: [12] `inferior-of` [14], trigger: [13][10].

Figure 1 shows the XML correspondent notations. For simplicity, only the `id` and the `lemma` attributes of the word elements `W` are shown. The `REFERENTIAL` elements, not marking boundaries, are indicated here as stand-off.

```
<ENTITY ID="E8" TYPE="PERSON">
 <W id="28" LEMMA="Marcus">Marcus</W>
 <W id="29" LEMMA="Vinicius">Vinicius</W>
</ENTITY>
<W id="30" LEMMA="fi">era</W>
<KINSHIP ID="KIN57" FROM="E12" TO="E11"
TRIGGER="31" TYPE="child-of">
 <ENTITY ID="E12" TYPE="PERSON">
  <W id="31" LEMMA="fiu">fiul</W>
  <KINSHIP ID="KIN53" FROM="E11" TO="E10"
TRIGGER="32" TYPE="sibling-of">
   <ENTITY ID="E11" TYPE="PERSON">
    <W id="32" LEMMA="soră">surorii</W>
    <ENTITY ID="E10" TYPE="PERSON">
     <W id="33" LEMMA="său">sale</W>
    </ENTITY>
    <W id="34" LEMMA="mai">mai</W>
    <W id="35" LEMMA="mare">mari</W>
   </ENTITY>
  </KINSHIP>
 </ENTITY>
</KINSHIP>
<W id="36" LEMMA=",">,</W>
<KINSHIP ID="KIN59" FROM="E13" TO="E15"
TRIGGER="44" TYPE="spouse-of">
 <ENTITY ID="E13" TYPE="PERSON">
  <W id="37" LEMMA="care">care</W>
 </ENTITY>
 <W id="38" LEMMA=",">,</W>
 <W id="39" LEMMA="cu">cu</W>
 <W id="40" LEMMA="an">ani</W>
 <W id="41" LEMMA="în_urmă">în urmă</W>
 <W id="42" LEMMA=",">,</W>
 <W id="43" LEMMA="sine">se</W>
 <W id="44" LEMMA="căsători">căsătorise</W>
 <W id="45" LEMMA="cu">cu</W>
 <KINSHIP ID="KIN61" FROM="E15" TO="E14"
TRIGGER="46" TYPE="parent-of">
  <ENTITY ID="E15" TYPE="PERSON">
   <W id="46" LEMMA="tată">tatăl</W>
    <ENTITY ID="E14" TYPE="PERSON">
     <W id="47" LEMMA="acesta">acestuia</W>
    </ENTITY>
   </ENTITY>
 </KINSHIP>
</KINSHIP>
<SOCIAL ID="SOC9" FROM="E17" TO="E16" TRIGGER="49"
TYPE="inferior-of">
 <ENTITY ID="E17" TYPE="PERSON">
  <W id="49" LEMMA="consul">consul</W>
  <W id="50" LEMMA="pe">pe</W>
  <W id="51" LEMMA="vreme">vremea</W>
  <W id="52" LEMMA="el">lui</W>
  <ENTITY ID="E16" TYPE="PERSON">
   <W id="53" LEMMA="Tiberiu">Tiberiu</W>
  </ENTITY>
 </ENTITY>
</SOCIAL>
<W id="54" LEMMA=".">.</W>

<REFERENTIAL ID="REF37" FROM="E12" TO="E8"
TYPE="coref" /REFERENTIAL>
<REFERENTIAL ID="REF38" FROM="E13" TO="E11"
TYPE="coref" /REFERENTIAL>
<REFERENTIAL ID="REF39" FROM="E14" TO="E8"
TYPE="coref" /REFERENTIAL>
<REFERENTIAL ID="REF40" FROM="E17" TO="E15"
TYPE="class-of" /REFERENTIAL>
```

Fig. 1. Example of an annotation

IV. A CORPUS INCORPORATING ENTITY LINKS

A. Building the corpus

A group of master students in Computational Linguistics, first year, annotated the corpus in a collaborative work. Each of them received approximately 10 pages of text from the Romanian version of Henryk Sienkiewicz's "Quo Vadis". The XML conventions and the annotation rules have been established in the group along a number of weekly debates that lasted more than a couple of months, by discussing many examples. Before starting the manual annotation activity, the text of the whole book was passed through an initial NLP chain which included a tokeniser, a POS-tagger and a lemmatiser[11]. The students were instructed to mark in a first step the entities and only in a second step the relations. The annotation tool used was PALinkA[12].

A number of limitations have been set, with the intention to reduce the complexity of the task. Here are they:

- We did not annotate negated relations. For instance, no relation is marked in case the verb linking the subject to the predicative noun is negated: 1:[*Ligia*] *nu poate să devină* 2:[*amanta nimănui*]. <1:[*Lygia*] *could not become* 2:[*the concubine of any man*]> => [2] and [1] are not linked by a (negated) coreferential relation; 1:[*Vinicius*]*... Nu eşti* 2:[*un oarecare*] *şi nu ai* 3:[*un chip de rând*]. <1:[*Stranger*]*, thou seemest no* 2:[*evil man*] *nor* 3:[*foolish*].> => no relation is marked between [2] and [1], nor between [3] and [1].
- Characterisations addressing subjects, expressed by predicative nouns, are not marked. Example: *Şi, ca un cunoscător, înţelese că* 1:[*este*; TYPE=INCLUDED] 2:[*o fiinţă deosebită*]. <*... and as a judge he understood that in* 1:[*her*] *there was* 2:[*something uncommon*]> => no relation is marked between [2] and [1].
- In this corpus, no interpreted relations, besides coreferential are marked yet. For instance, in the following span, the `name-of-interpret` relation is not marked between [2] and [1]: *Petronius... care*

[10] Note that in the English variant, the social relation is not that evident as in the Romanian variant.

[11] Web services of the NLP-Group@UAIC-FII, at http://nlptools.info.uaic.ro/.

[12] http://clg.wlv.ac.uk/projects/PALinkA, author Constantin Orasan.

simțea că pe statuia 1:[*acestei fete*] *trebuia scris* 2:[*"Primăvara"*]. <*...who felt that beneath a statue of* 1:[*that maiden*] *one might write* 2:[*"Spring."*]>.

- Care was taken where lexicals used as triggers had different senses than those implied by the relation. For instance, in the context *La botezul sfânt mi s-a dat numele de Urban, părinte. <At holy baptism, father, the name Urban was given me.>*, the word *părinte <father>* is not a trigger for a `parent-of` type of kinship relation, because its sense here is *priest*.
- In case of coreferentiality (the anaphoric relations `coref`), the annotators were instructed to mark no more than *N*-1 relations in a coreferential cluster of *N* REs, therefore the minimum number of relations which are necessary to recover the complete cluster by symmetry and transitivity. This allows for exactly one initial member (first mention) and exactly one coreference link for each succeeding mentions of the same entity. Also, any member of the cluster can be chosen as antecedent of any of the but-first members of the cluster[13]. In the following example: *Se repezi la* 1:[*Petru*] *și, luându*-2:[*i*] 3:[*mâinile*]*, începu să* 4:[*i*] 5:[*le*] *sărute. <...seized* 3:[*the hand of* 1:[*the old Galilean*]]*, and pressed* 5:[*it*] *in gratitude to his lips.*[14]*>* => [2] `coref` [1], [3] `part-of` [1] (or [3] `part-of` [2]), [4] `coref` [1] (or [4] `coref` [2]), [5] `coref` [3]. As such, marking [5] `part-of` [1] is superfluous, because it results, by transitivity, from [5] `coref` [3] and [3] `part-of` [1]. In the following span 1:[*Numele de familie al lui* 2:[*Ion*]] *este* 3:[*Rădulescu*]. <1:[*The family name of* 2:[*John*]] *is* 3:[*Rădulescu*].> the anaphoric relations can be annotated two ways: [1] `name-of` [2] (and not [2] `has-name` [1], because the convention in case of anaphoric relations between nested REs, as stated in Section III.B, is from the outer RE towards the inner RE), plus [3] `coref` [1]. An equivalent notation is [3] `name-of` [2], [3] `coref` [1].

In its present shape the corpus covers 1,500 sentences and includes: 3663 referential expressions, 2045 anaphoric links, 39 affective relations, 21 kinship relations and 15 social relations. In the training sessions, the students were instructed to give more importance to the correctness of annotations then to their completion. Moreover, the annotations received a second look through by the second author. In this initial phase of the research, we will mainly observe the precision of an automatic parser then its recall.

[13] This is true even in cases when the chain includes a proper noun followed by a number of pronouns and the coreferentiality is decided between some pairs of pronouns, provided (at least) one of them is also found coreferential with the initial proper noun. Let's note that it is important to know that a series of pronouns refer the same entity. The proper identity of these mentions, once decided for one of them, is transferred by symmetry and transitivity to all members of the cluster.

[14] In the English equivalent, two mentions of Peter ([2] and [4]) are missing).

V. Conclusions

Deep understanding of texts is a challenge that is still far from being accomplished. Meanwhile, emergent domains approach sub-tasks of this giant enterprise, among them: word sense disambiguation, syntactic parsing, discourse parsing, metaphor interpretation, textual entailment, semantic role labelling, paraphrase interpretation and generation, and entity linking. The task is so difficult because it necessitates large corpora annotated with expert knowledge, at different layers of interpretation of language. Entity linking or the task of recognising relations linking mentions of entities in texts, is perhaps, by its ambitions, the most close to this distant goal. A text contains static descriptions about entities, events relying them, general statements about world or about a small part of it. The segmentation that is natural in language, as given by syntax (words, clauses, sentences) offer the ingredients of the language structure, but no one knows yet in what terms a representation should be stated. In principle, such a representation should be stable to variations which do not confuse the content and to any source languages.

The research described in this paper relates to the aspect of inventing a representation for deep text understanding that would help the technologies of entity linking. It focuses on representing relations between characters in fiction texts. Criteria to annotate referential expressions that realise entities have been established. Then we focussed on four types of relations between entities of type person and proposed conventions for annotating them in XML.

The research showed that the human annotators, although accomplishing for the first time such a task, generally responded well to our training in the attempt to annotate unambiguously relations between entities. In order to ease the annotation process and to reduce at minimum possibilities of variation due to personal interpretations, a number of conventions were established. Among such constraints were, for instance, those intended to detect the exact span of referential expressions, the borders delimiting relations, the criteria to decide between a relation and its inverse in case of non-symmetrical relations, when and how to mark a null pronoun, etc.

Further research will concentrate on two directions. First the corpus must be enlarged, its accuracy – augmented and its completion – scrutinised. One activity we have to develop in the near future will be to monitor parallel annotations done by different subjects and measure the inter-annotator agreement. When accomplished, this direction will produce a gold corpus that will sustain the second line of research: the development of entity linking algorithms trained and, afterwards, validated on the corpus.

Acknowledgment

We are grateful to the class of first year master students in Computational Linguistics, from the Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași, who, during the second term of the university year 2012-2013, realised the "Quo Vadis" corpus described in this paper.

REFERENCES

[1] R.C. Bunescu and M. Pasca, "Using encyclopedic knowledge for named entity disambiguation". European Chapter of the Assocation for Computational Linguistics (EACL), 2006.

[2] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data". Empirical Methods in Natural Language Processing (EMNLP), 2007.

[3] D. Rao, P. McNamee, and M. Dredze, "Entity Linking: Finding Extracted Entities in a Knowledge Base", in Thierry Poibeau, Horacio Saggion, Jakub Piskorski and RomanYangarber (eds.) Multisource, Multilingual Information Extraction and Summarization, Springer Lecture Notes in Computer Science, Berlin, Heidelberg, 2012.

[4] A. Carlson, J. Betteridge, E.R. Hruschka Jr. and T.M. Mitchell, "Coupling Semi-Supervised Learning of Categories and Relations". In Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, 2009.

[5] T. Mohamed, E.R. Hruschka Jr. and T.M. Mitchell, "Discovering Relations between Noun Categories". In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.

[6] P.P. Talukdar, D.T. Wijaya and T.M. Mitchell, "Acquiring Temporal Constraints between Relations". In Proceedings of the Conference on Information and Knowledge Management (CIKM), 2012.

[7] P. McNamee, J. Mayfield, T. Finin, T. Oates, D. Lawrie, T. Xu, and D. W. Oard, "KELVIN: a tool for automated knowledge base construction". In Proceedings of the NAACL HLT 2013 Demonstration Session, pp. 32-35, Atlanta, 10-12 June 2013.

[8] E. Boschee, R. Weischedel, and A. Zamanian. "Automatic information extraction". In Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, VA, pages 2–4, 2005.

[9] A. Bagga and B. Balwdin, "Entity-based cross-document coreferencing using the Vector Space Model", in COLING '98 Proceedings, vol. 1, 1998.

[10] A. Bagga and B. Balwdin, "Cross-document event coreference: annotations, experiments, and observations", CorefApp '99 Proceedings, 1999.

[11] H. Saggion, "SHEF: semantic tagging and summarization techniques applied to cross-document coreference, SEMEVLA '07 Proceedings, 2007.

[12] S. Singh, A. Subramanya, F. Pereira and A. McCallum, "Large-scale cross-document coreference using distributed inference and hierarchical models", HLT '11 Proceedings, vol. 1, 2011.

[13] D. Cristea and G.E. Dima, "An integrating framework for anaphora resolution". Information Science and Technology, Romanian Academy Publishing House, Bucharest, vol. 4, no. 3-4, p 273-291, 2001.

[14] C. Orăsan, D. Cristea, R. Mitkov, A.H. Branco, "Anaphora Resolution Exercise - An Overview". In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Marrakech, 26 May - 1 June 2008.

[15] D. Cristea, O. Postolache, "How to deal with wicked anaphora". In Antonio Branco, Tony McEnery, Ruslan Mitkov (eds.): Anaphora Processing: Linguistic, Cognitive and Computational Modelling, Benjamin Publishing Books, 2005.

[16] D. Anechitei, D. Cristea, I. Dimosthenis, E. Ignat, D. Karagiozov, S. Koeva, M. Kopeć, C. Vertan, "Summarizing Short Texts Through a Discourse-Centered Approach in a Multilingual Context". In Neustein, A., Markowitz, J.A. (eds.), Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems. Springer Verlag, Heidelberg/New York, 2013.

[17] R. Simionescu. "Romanian Deep Noun Phrase Chunking Using Graphical Grammar Studio". In Proceedings of the ConsILR conference 2011-2012.

[18] M. Colhon, "Syntactic Translation Patterns from a Parallel Treebank", Proceedings of the First Workshop on Computational Linguistics of Balkan Languages (CLoBL 2012), The 5th Balkan Conference in Informatics – BCI 2012, Novi Sad, Serbia, September 16-20, ISBN 978-86-7031-200-5, pp. 85-88, 2012.

[19] D. Anechitei, "Multilingual Discourse Processing". Disertation thesis, "Alexandru Ioan Cuza" University of Iași, Department of computer Science, 2012.

[20] A. Iftene, "Textual Entailment", PhD thesis, "Alexandru Ioan Cuza" University of Iași, Department of Computer Science, Iași, 2009.

[21] A. Moruz, "Predication Driven Textual Entailment", PhD thesis, "Alexandru Ioan Cuza" University of Iași, Department of Computer Science, Iași, 2011.

[22] D. Trandabăț, "Natural Language Processing Using Semantic Roles", PhD thesis, "Alexandru Ioan Cuza" University of Iași, Department of Computer Science, Iași, 2010.

[23] G. Cărăușu, "Processing Spatial Relations In Old Texts And Their Transposition On Modern Maps" (in Romanian: "Prelucrarea expresiilor spațiale în textele vechi pentru realizarea echivalărilor topografice în hărțile actuale"), graduation thesis, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași, 2011.

[24] A.M. Ciucanu, "Iter in Chinam – Reconstructing the Journey of Milescu Spătarul from Russia to China" (in Romanian "Iter in Chinam – Reconstituirea traseului lui Milescu Spătarul din Rusia până în China"), graduation thesis, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași, 2011.

[25] D. Ferruci, A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment", Natural Language Engineering 10, No. 3-4, pp. 327-348, 2004.

[26] K.Yoshinobu, W.A. Baumgartner Jr., L. McCrohon, S. Ananiadou, K. B. Cohen, L. Hunter and J. Tsujii, "U-Compare: share and compare text mining tools with UIMA", Bioinformatics. 25(15), pp. 1997-1998, 2009.

[27] B.J. Grosz, A. Joshi, S. Weinstein, "Centering – a framework for modelling the local coherence of discourse". In: Computation Linguistics, 12(2), June 1995.