Semantic Frames in Romanian Natural Language Processing Systems

Diana Maria Trandabăț
Faculty of Computer Science, "Al.I.Cuza" University of Iaşi &
Institute for Computer Science, Romanian Academy, Iaşi Branch
16, Gen. Berthelot, 700483-Iaşi, Romania
dtrandabat@info.uaic.ro

Abstract

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. I thus propose to create a semantic frame database for Romanian, similar to the FrameNet database. Since creating language resources demands many temporal, financial and human resources, a possible solution could be the import of standardized annotation of a resource developed for a specific language to other languages. This paper presents such a method for the importing of the FrameNet annotation from English to Romanian.

1 Introduction

The realization of human-computer interaction in natural language represents a major challenge in the context of aligning Romania to existing technologies.

The proposed project aims to introduce the semantic frames and contexts, which define a concept's sense according to its facultative or mandatory valences (Baker and Fillmore, 1998), to Romanian NLP systems. The behavior of the Romanian clauses – mainly the verbal group, around which all the other sentence complements gravitates in a (more or less) specific order – has been closely debated in the last years (Irimia, 1997; Dobrovie-Sorin, 1994; Monachesi, 1998; Barbu, 1999), creating a proper frame for the introduction of semantic roles.

This paper presents the steps considered for the achievement of this project. Thus, Section 2 gives a very brief description of the frame semantics, and Section 3 presents the realization of a semantic structures database for the Romanian language, similar to those existing for English, German, or Spanish, containing detailed information about the relations between the semantic meaning and the syntax of the words. In the last section, some possible applications of the detection of semantic roles to written and spoken texts are mentioned (question answering systems, summarization systems, prosody prediction systems), before drawing some final conclusions.

2 Frame Semantics

The FrameNet (FN) lexical-semantic resource is based on the principles of Frame Semantics (FS). From FS point of view, the semantic/syntactic features of "*predicational words*"[1] (Curteanu, 2003-2004) are defined in a particular semantic frame. The sentences are schematic representations of different situations, including different participants, objects or other conceptual roles. Being a linguistically transposed experience, a sentence represents an event scenario that is structured around a semantic head. The meaning of this head

[1] Words, mostly verbs, but also several nouns and adjectives, bearing a *predicational feature*, viz. demanding a specific semantic argument structure in order to complete their meaning.

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 29–32,
Rochester, April 2007. ©2007 Association for Computational Linguistics

can be understood only by expressing the core frame elements and can, optionally, be enriched with other semantic features, by expressing some non-core frame elements.

Fillmore (1968) divides 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.

The Case Notions are representations at a semantic level of the lexical arguments. This inventory of cases comprises universal concepts, possible innate, sufficient for the classification of the verbs of a language and reusable in all languages. The list of Fillmore Cases, which will be considered for the project, includes: Agent, Instrument, Dative, Experiencer, Locative, Object, etc.

3 A Parallel Romanian/English FrameNet Using Annotation Import

The first step in the realization of the Romanian corpus of annotated semantic frames was the manual translation 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. This frame was selected due to its rich frame to frame relations (Inheritance – *Change_of_consistency, Process_start,* etc., Subframe - *Change_of_state_scenario* and Using - *Process_end*). After the selection of the clauses and their translation, the Romanian sentences were aligned with the English ones using the aligner developed by the Institute of Research in Artificial Intelligence (Tufiş et al., 2005). The next step was the automatic import of the English annotation, followed by a manual verification, a detection of the mismatching cases and an optimization process which, based on inference rules, corrects the automatic annotation.

3.1 Automatic annotation import

The intuition behind the importing program (Trandabăț et al., 2005) 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. The surface realization is realized according to each language syntactic constraints.

The automatic importing program 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. The automatic import is manually checked in order to establish the method efficiency.

3.2 The algorithm

The starting point for the German, Japanese and Spanish FN creation was the manual annotation at FE level of existing corpora for each language. For Romanian, I propose creating a corpus of semantic roles starting from the translation of (a part of) the English corpus of annotated sentences (see Figure 1).

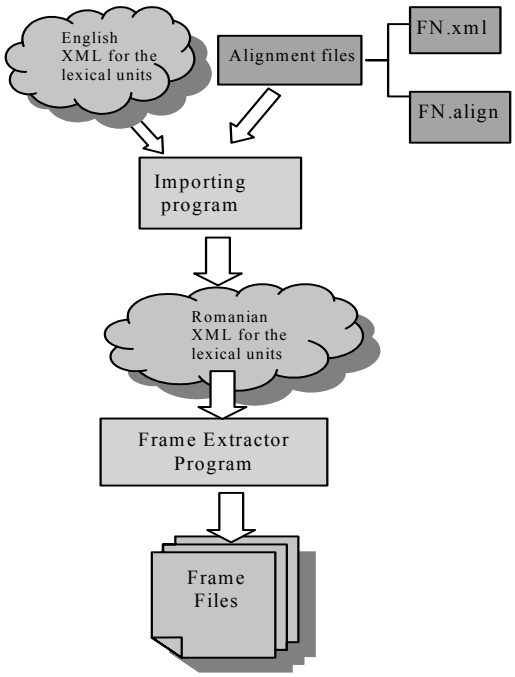

Figure 1. The architecture of the importing program

Using the XML files of the annotated English sentences, and the alignment files where each English word is linked to its corresponding Romanian translation, I automatically created a set of XML files containing a corpus of FE annotated sentences for the Romanian language. An example of imported annotation for the English lexical unit „occur" is presented in figure 2.

```
<annotationSet ID="1" status="AUTOMATIC">
 <layers>
  <layer ID="6375447" name="FE">
   <labels>
    <label name="Event" ID="19909459"
cDate="june 2006" start="0" end="9" />
    <label name="Time" ID="19909462"
cDate="june 2006" start="20" end="59" />
    <label name="Place" ID="19909465"
cDate="june 2006" start="61" end="101" />
   </labels>
  </layer>
..........
  <layer ID="6375452" name="Target">
   <labels>
    <label name="Target" ID="19905041"
cDate="june 2006" start="11" end="18" />
   </labels>
  </layer>
  <layer ID="6375453" name="Verb" />
 </layers>
 <sentence ID="671" aPos="103724676">
  <text>Incidentul a apărut după o dispută
între individ şi personal la o filială a
Băncii Irlandeze din Cahir .
  </text>
 </sentence>
</annotationSet>
```

Figure 2. Example of annotation set for English

The <*annotationSet*> tag indicates that a new sentence is annotated. Inside this tag, the <layers> tag sets the annotation layer (FE - Frame Element, GF - Grammatical Function or PT - Phrase Type) and the <*sentence*> tag encloses the text. The labels are applied to the words in <*text*>, indexed by their character. For example, the tag:

```
<label name="Event" ID="19909459"
cDate="june 2006" start="0" end="9"
/>
```

indicates that the *Event* frame element is starting with the first character of the sentence an stops at the 9th, meaning that the *Event* FE corresponds to „*Incidentul*" (en. *The incident*).

The general algorithm of the automatic importing program focuses on:

- reading of the input XML files;
- labeling of each English word with the corresponding semantic role (FE)
- converting the character indexes into a word level annotation;
- mapping the English words with the aligned Romanian correspondences, hence with the respective semantic role;
- writing an output XML file containing the Romanian annotated corpus.

For example, the lexical unit "occur.v" will appear in English and Romanian annotated as:

[Incidental]$_{Event}$ A APĂRUT [după o dispută între individ şi persona]$_{time/cause}$ [la o filială a Băncii Irlandeze din Cahir]$_{Place}$.
The incident]$_{Event}$ OCCURRED [after a dispute between the man and staff]$_{time/cause}$ [at a branch of the Bank of Ireland in Cahir]$_{Place}$

3.3 Optimization

My initial experiment has involved the translation of approx. 1000 sentences from English FN. The translations have been realized by professional translators, so the errors propagated in the corpus should be minimal. 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 alignment process was performed with the aligner developed by the Institute of Research in Artificial Intelligence (Tufiş, 2005), which is considered to have a precision of 87.17% and a recall of 70.25%. However, the aligner results were manually validated before entering the annotation import program.

The assessment of the correctitude of the obtained Romanian corpus is preformed manually. The first results of the annotation import show an overall accuracy of approx. 80%. The validation focuses on detecting the cases where the import has failed, trying to discover if the problems are due to the translation 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 were cases where the FEs are expressed in English, but are implicit in the Romanian translation, as in:

[Blood]$_{Undergoer}$ had CONGEALED [thickly]$_{Manner}$ [on the end of the smashed fibula]$_{Place}$.

[Sângele]$_{Undergoer}$ se ÎNGROŞĂ [spre capătul fibulei zdrobite]$_{Place}$.

or not-expressed in English, but expressed in Romanian, as the *Protagonist* role in :

QUIT [smoking]$_{Process}$.
LĂSAŢI-[vă]$_{Protagonist}$ [de fumat]$_{Process}$.

The frame generation program based on the generated Romanian corpus is currently under development.

4 Conclusions

In this paper, I have presented a fast method for the realization of a Romanian corpus annotated with semantic frame relations. The main purpose of creating a quick semantic annotated database is using it as training corpus for automatic labeled semantic frames detection. Nowadays, expensive linguistic resources demanding a lot of time, money and human resources are created for different languages. After their utility is proved, those resources begin to be imported to other languages (see for instance the MultiSemCor project[2]). In this context, the realization of a Romanian FN is a challenging project in the frame of Romance FN.

The import method was preferred to the 'classical' creation by hand of a manually annotated corpus because of its possible automation. I investigate currently the possibility of using a translation engine for the most time consuming task, namely the translation of the English sentences. The project will be further developed by adding to the automatic import program rules discovered though the analysis of the mismatching cases.

The lack of semantic information was very obvious while working on the QA@CLEF competition[3] (Question Answering task within the Cross Language Evaluation Forum Competition) last year (Puşcaşu et al., 2006); having the semantic frames database (thus a semi-automatic role labeling system) can improve the precision of selecting an appropriate snippet for the desired answer, not to mention also the benefits for answer generation. Another application of the semantic frames I am interested in is prosody prediction. Within the Institute of Computer Science, I have begin to work at a syntax-prosody interface for Romanian based on FDG trees of sentences and other syntactical information to discover the phonological entities underlying the written text and the topic/focus articulation. The algorithm for finding sentence focus uses the semantic roles as a main component.

References

Baker, C., Fillmore, Ch., Lowe, J., *The Berkeley FrameNet project*, in Proceedings of the COLING-ACL, Montreal, Canada, 1998

Barbu, A-M, *The Verbal Complex*. Linguistic Studies and Enquires, L, no.1, Bucharest, p. 39-84 (In Romanian). 1999

Curteanu, N.: *Contrastive Meanings of the Terms "Predicative" and "Predicational" in Various Linguistic Theories (I, II)*. Computer Science Journal of Moldova (R. Moldova), Vol. 11, No. 4, 2003 (I); Vol. 12, No. 1, 2004 (II)

Curteanu, N., Trandabăţ, D., Moruz, M.: *Substructures of the (Romanian) Predicate and Predication Using FX-bar Projection Functions on the Syntactic Interface,* in Proc. of the 4th European Conference on Intelligent Systems and Technologies - ECIT2006, Iaşi, Romania, 2006.

Dobrovie-Sorin, C, *The syntax of Romanian. Comparative Studies.* Berlin: Mouton de Gruyter, 1994

Fillmore, Ch., *The case for case*; in Bach and Harms (Eds.), Universals in Linguistic Theory, Ed. Holt, Rinehart, and Winston, New York, 1968

Husarciuc M, Trandabăţ D., Lupu M., *Inferring Rules in Importing Semantic Frames from English FrameNet onto Romanian FrameNet,* 1st ROMANCE FrameNet Workshop, EUROLAN, Cluj, Romania 2005

Irimia, D. *The Morphosyntax of the Romanian* Verb. Ed. of the "Al. I. Cuza" Iaşi University (in Romanian). 1997

Monachesi, P., *The Morphosyntax of Romanian Cliticization*. in: P. A. Coppen et al. (Eds.), Proceedings of Computational Linguistics in The Netherlands, pp. 99-118, Amsterdam-Atlanta: Rodopi. 1998

Puşcaşu, G., Iftene, A., Pistol, I., Trandabăţ, D., Tufiş, D., Ceauşu, A., Ştefănescu, D., Ion, R., Orăşan, C., Dornescu, I., Moruz, A., Cristea, D., *Developing a Question Answering System for the Romanian-English Track at CLEF 2006*, CLEF 2006 Workshop, Alicante, Spain, to be published in LNCS

Trandabăţ, D., Husarciuc, M., Lupu, M., *Towards an automatic import of English FrameNet frames into the Romanian language*, 1st ROMANCE FrameNet Workshop, EUROLAN, Cluj, Romania, 2005

Tufiş, D., Ion R., Ceauşu, Al., Stefănescu, D., *Combined Aligners* in Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond", Ann Arbor, Michigan, June, 2005

[2] http://multisemcor.itc.it/
[3] http://clef-qa.itc.it/2006bis/CLEF-2006.html