Towards automatic cross-lingual transfer of semantic annotation

Diana Trandabăț

Faculty of Computer Science, University "Al. I. Cuza" Iași
14 Alexandru Lăpușneanu
700057 Iași
Romania
dtrandabat@info.uaic.ro

ABSTRACT. *In order to develop a semantic labeling system, the most common methods use supervised learning from an annotated corpus. What if we have short deadlines and limited human and financial possibilities that prevent us from building such a training corpus for our language? If such a corpus already exists for any other language, this paper proposes a method to automatically import the existing corpus for the language we need. The transfer method is based on translating the existing corpus (or using annotated versions of existing parallel texts), aligning it at word level, and applying a set of mapping functions to import the annotation from one language to another. An import validation interface is also offered for the manual validation of the resulted resource. As an example, the case of semantic role import from the English FrameNet to Romanian is discussed.*

RÉSUMÉ. *Afin de développer un système d'étiquetage sémantique automatique, les méthodes les plus fréquentes utilisent l'apprentissage supervisé à partir d'un corpus annoté. Et si on a des délais courts et des possibilités humaines et financières limitées, qui nous empêchent de construire un tel corpus d'apprentissage pour la langue de notre choix? Si un tel corpus existe déjà pour une autre langue, cet article propose une méthode pour importer automatiquement le corpus existant dans la langue où nous le nécessitons. La méthode de transfert présentée dans cet article est basée sur la traduction du corpus existant (ou l'utilisation d'une version parallèle annotée du texte), l'alignement au niveau du mot des deux versions de texte, et l'application d'un set de fonctions de mappage pour importer l'annotation d'une langue à l'autre. Une interface de validation de l'import est également offerte pour la validation manuelle de la ressource obtenue. A titre d'exemple, le cas de l'import des rôles sémantiques de la ressource anglaise FrameNet vers le roumain est discuté.*

KEY WORDS: *Natural language processing, semantic frames, semantic role labeling system, semantic annotation import, computer-aided text annotation.*

MOTS-CLÉS: *Traitement du langage naturel, cadres sémantiques, système d'étiquetage des rôles sémantiques, annotation semi-supervisée.*

1. Introduction

A key concern in the Information Retrieval domain is the identification of the mechanism that allows the attachment of meaning to larger chunks of text, including the study of sense and denotative references, argument structures, semantic roles, discourse analysis, and the linking of all of these to syntax. The area this paper intends to cover is semantic role analysis, intending to identify the role each entity plays in different events.

The main question this paper 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 (Backer et el., 1998), and later for German (Erk et al., 2003), Spanish (Subirats-Ruggeberg et al., 2003) 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 investigate in this paper the success rate of the transfer of semantic role annotation from English to Romanian. After this resource is built, an automatic role labeler can be created using supervised learning to annotate raw text with the semantic role information, for usage in natural language processing tools.

This paper presents this semantic transfer[1] method, being organized as follows: Section 2 presents the state of the art in semantic role annotation and several existing resources, Section 3 describes in detail the transfer method we used, with exemplifications for the import of semantic annotation from English to Romanian, and Section 4 discussed the results and further development of the proposed method.

2. State of the art

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. In the last decades, hand-tagged corpora that encode semantic role information for the English language were developed, adopting different sets of roles. **VerbNet**[2] extends Levin's classes (Levin et el., 2005) by adding an abstract representation of the syntactic frames for each class. **FrameNet**[3] produced a lexicon containing very detailed information about the syntax - semantics relations of the English predicate words (verbs, nouns and adjectives). The **Proposition Bank** project[4] added a layer of predicate-argument information to the syntactic structures of the Penn Treebank (Palmer et al., 2005).

---

[1] Throughout this paper, the terms of semantic annotation „transfer“ and semantic annotation „import“ are interchangeably used, referring to the same process.
[2] VerbNet web address: http://verbs.colorado.edu/ kipper/verbnet.html
[3] FrameNet web address: http://framenet.icsi.berkeley.edu
[4] PropBank web address: http://www.cs.rochester.edu/ gildea/PropBank/Sort

In building semantic resources for languages other than English, two methods can be adopted, as extracted from the methods used by (Vossen, 1999) for mapping multilingual resources: 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. With the *Expand approach*, a resource for one language, which is regarded as stable at that time, is transferred to another language. 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. The semantic resources built along the FrameNet model (the German, Spanish or Japaneese FrameNet) use the *Merge* approach, and allow for cross-lingual relations at many levels, including those of frames, lexical units, lemmas, or even word forms and annotated sentences. 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. The semantic annotation transfer method proposed in this paper differs by adopting the *Expand* method.

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. 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), Johansson and Nugues (2005) and Tonelli and Pianta, 2008. This paper presents the formalization of the method used for the transfer of semantic annotation from English FrameNet to Romanian based on sentences aligned at word level.

3. Automatic transfer of semantic annotation

The Semantic Roles Import program is based on the assumption that, if a word is part of a specific role in English, it will be part of the same role type in Romanian (Trandabat, 2007). As an example, consider:

EN: ...until [Craig]$_{\text{Entity}}$ [becomes]$_{\text{TARGET}}$ [available]$_{\text{Finalstate}}$ [in 1994]$_{\text{Time}}$.
RO :...până când [Craig]$_{\text{Entity}}$ [va deveni]$_{\text{TARGET}}$ [disponibil]$_{\text{Finalstate}}$ [în 1994]$_{\text{Time}}$.
FR :… jusqu'à ce que [Craig]$_{\text{Entity}}$ [devient]$_{\text{TARGET}}$ [disponible]$_{\text{Finalstate}}$ [en 1994]$_{\text{Time}}$.

The import program needs as prerequisites translation of the annotated English sentences into the desired language (for this exemplification Romanian), which can be performed either by a human translator or by Google translate service. The next step is the alignment of the original and translated versions of the sentences using Giza++. These files enter into the semantic roles import program. After the import of the semantic annotation, an interface can be used to perform manual validation of

the imported corpus, in order to detect the mismatching cases. Another module, in work at this moment, will offer an optimization process which, based on inference rules extracted from the user corrected transfers, corrects the automatic annotation.

The transfer algorithm considers the import as a sequential labeling problem, with a B-I-O encoding. In the B-I-O representation, a word is characterized a being at the **B**eginning, **I**nside or **O**utside of a sequence to be analyzed, in our case, of a semantic frame (Trandabat 2010). The automatically 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. 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} , . . . , w_{en}) and the Romanian translation can be considered S_ro = (w_{r1} , w_{r2} , . . . , w_{rm}), thus the alignment function is

$$\text{Align}(w_{ei}) : \text{S\_en} \cup \{\emptyset\} \rightarrow 2^{\text{S\_ro}} \cup \{\emptyset\}, \quad (1)$$

where $2^{\text{S\_ro}}$ is the set of all possible parts of the Romanian sentence S_ro.

We assign to the English words their annotated frames through Frame(w_{ei}):

$$\forall w_{ei} \in \text{S\_en}, \text{Frame}(w_{ei}) : \{\text{S\_en} \cup \emptyset\} \rightarrow \{\text{B\_F}_i , \text{I\_F}_i , \text{O\_F}_i \}, \quad (2)$$

with $F_i \in \{\cup$ FrameNetRoles, NO-Frame$\}$ and Frame($\emptyset$) = $\{\emptyset\}$.

The assignment of 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 Align mapping function, $\forall w_{ei} \in$ S_en, Align(w_{ei}) = $W^i_R \subset$ S_ro $\cup \{\emptyset\}$, we have the following transfer cases:

one-to-zero alignment: $\exists!\ w_{ei} \in$ S_en, such that Align(w_{ei}) = $\emptyset$, meaning that an English word has no Romanian correspondent, i.e. no transfer function is needed;

one-to-one alignment, when $\exists!\ w_{ei} \in$ S_en and $\exists!\ w_{rj} \in$ I S_ro, such that Align(w_{ei}) = w_{rj} , with $| W^i_R | = 1$, meaning that an English word has one and only one Romanian correspondent, and the transfer function Frame(w_{rj}) = Frame(w_{ei});

one-to-many alignment, when $\exists!\ w_{ei} \in$ S_en and a subset of Romanian words w_{rj} . . . $w_{rl} \in 2^{\text{S\_ro}}$, such that Align(w_{ei}) = w_{rj} . . . w_{rl} , with $| W^i_R | > 1$, meaning that an English word is translated with more than one Romanian words, in which case the transfer function becomes:

$$\textit{Frame}(w_{r_j}) = \begin{cases} \textit{Frame}(w_{e_i}) \text{ if } \textit{Frame}(w_{e_i}) = \{B\_F_i\} \text{ and we are at the first Romanian word} \\ \qquad\qquad \text{in the alignment set } \textit{Align}(w_{e_i}) \\ \textit{Frame}(w_{e_i}) \text{ if } \textit{Frame}(w_{e_i}) \in \{\text{I\_F}_i, \text{O\_F}_i\} \\ \text{I\_F}_i \qquad \text{otherwise} \end{cases}$$

many-to-one alignment, when there exist a subset w_{ei} . . . $w_{ek} \in 2^{\text{S\_en}}$ and $\exists!\ w_{rj} \in$ S_ro, such that Align(w_{ei} , . . . w_{ek}) = w_{rj} , with $| W^i_R | = 1$, meaning that several English words are translated with the same Romanian word. In this case, the transfer function becomes:

$$Frame(w_{r_j}) = \begin{cases} F_i & \text{if } F_i = TARGET \text{ or } F_k = NO\_Frame \\ F_k & \text{otherwise} \end{cases}$$

many-to-zero alignment, when there exist a subset $w_{ei} \ldots w_{ek} \in 2^{S\_en}$, such that $Align(w_{ei} \ldots w_{ek}) = \emptyset$, $| W^i_R | = 0$, 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 exist a subset $w_{ei} \ldots w_{ek} \in 2^{S\_en}$ and a subset $w_{rj} \ldots w_{rl} \in 2^{S\_ro}$, such that $Align(w_{ei} \ldots w_{ek}) = w_{rj} \ldots w_{rl}$ and $| W^i_R | > 1$. This case is mostly theoretical, since in transferring English annotation to Romanian, no such alignment was found.

zero-to-one alignment, when for $w_{ei} = \emptyset$ $\exists!\ w_{rj} \in S\_ro$, such that $Align(w_{ei}) = w_{rj}$, 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;

$$Frame(w_{r_j}) = \begin{cases} I\_F_i & \text{if } Frame(w_{r_{j-1}} = \text{B}\_F_i \text{ or } I\_F_i \text{ and } Frame(w_{r_{j+1}}) = I\_F_i \\ O\_NO - Frame & \text{otherwise} \end{cases}$$

zero-to-many alignment, when for $w_{ei} = \emptyset \in S\_en$, there exist a subset $w_{rj} \ldots w_{rl} \in 2^{S\_ro}$, such that $Align(w_{ei}) = w_{rj} \ldots w_{rl}$ and $| W^i_R | > 1$. This situation is reduced to zero-to-one alignment, applied for each Romanian word;

zero-to-zero alignment, when $w_{ei} = \emptyset \in S\_en$, $Align(w_{ei}) = \emptyset$. 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.

Exemplified transfer cases can be found in (Trandabăț 2010).

The assessment of the correctness of the obtained Romanian corpus is performed by comparing it to the validated semantic role frames for Romanian (Trandabat 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, are (1) the double annotation, (2) the existence of imbricate frame elements (FEs) or (3) the unexpressed semantic frames.

4. Discussion and further work

This paper has introduced 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. A further development of our method involves using more sophisticated transfers than between aligned words, such as aligned constituents or phrases.

The semantic roles encode important information that can improve natural language processing systems. Therefore, several interactions between semantic roles and different tasks of computational linguistics can be envisaged, such as summarizing systems or question answering based on semantic information.

Acknowledgements

The research presented in this paper was funded by the Sectorial Operational Programme for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944.

5. Bibliography/References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. "The Berkeley FrameNet project". In *Proc. COLING-ACL 1998*.

Hwa R., Resnik, Ph., Weinberg, A., Kolak O., „Evaluating translational correspondence using annotation projection". In *Proc. ACL 2002.*

R. Johansson and P. Nugues. "Using Parallel Corpora for Cross-language Projection of FrameNet Annotation". *In ROMANCE FrameNet Workshop at EUROLAN 2005*.

Erk, K., Kowalski, A., Pado, S., Pinkal, M., "Towards a resource, for lexical semantics: A large German corpus with extensive semantic annotation", in *Proc. ACL 2003*.

Levin, Beth and M. Rappaport Hovav. "Argument Realization". *Research Surveys in Linguistics Series*. Cambridge University Press, Cambridge, UK, 2005.

Palmer, Martha Daniel Gildea, and Paul Kingsbury. *The proposition bank: An annotated corpus of semantic roles*. Computational Linguistics, 31(1):71-106, 2005.

Subirats-Ruggeberg, Carlos and Miriam R. L. Petruck. "Surprise: Spanish FrameNet! In Workshop on Frame Semantics", In *Proc. Int. Congress of Linguists*, Prague, 2003.

Tonelli S., Pianta E., "Frame information transfer from English to Italian", in *Proc. LREC'08.*

Trandabat D. and Husarciuc M., "Romanian semantic role resource", in *Proc. LREC'08*.

Trandabat D. . "Semantic frames in Romanian natural language processing systems". In *Proc. NAACL-HLT 2007 Doctoral Consortium*, pp. 29-32.

Trandabăț D., *Natural Language Processing Using Semantic Frames*, PhD Thesis, University Al. I. Cuza Iasi, Romania.

P. Vossen. *EuroWordNet general document*. Version 3, Rapport, Univ. of Amsterdam, 1999.