SENSE-TAGGING OF ROMANIAN GLOSSES

CORINA HOLBAN, FELICIA CODIRLAȘU,
ANDREI MINCĂ, ȘTEFAN DIACONESCU[1]

[1] *Research and Development Department, SOFTWIN, Bucharest – Romania*
{cholban , fcodirlasu, aminca , sdiaconescu}@softwin.ro

Abstract

This work describes a research phase carried out in the project SenDiS ("General Word Sense Disambiguation System applied to Romanian and English Languages") that is the creation of a specific lexicon network obtained by manually tagging the tokens in a lexicon's glosses with their contextual meaning. Therefore, we present here: the annotation model, the tool used for manually tagging the glosses and some statistics following the annotation effort. Also, we give here the set of 'sense-tagging' principles derived from the preliminary analysis of glosses and enriched afterwards by the tagging experience.

Key words — knowledge-based Natural Language Processing, lexicon networks, Lesk algorithm, semantically annotated glosses, WordNet tagged glosses

1. Introduction

Supervised and especially unsupervised paradigms are the nowadays popular approaches to Natural Language Processing (NLP). Both exploit the increasing computing power of machines and also the exponential growth of digital information.

Sometimes, this process is accelerated by private initiatives of great effort. Consider only the Google Print Library Project, started in 2004, which was to digitize and make available approximately 15 million volumes within a decade. With this undertaking, Google also made the switch from rule-based NLP to statistical NLP and machine learning.

However, at the same time, carrying the Google Print Library Project sent an important message: the need of qualitative or structured data as input for better NLP. Whether the choice should be statistical NLP or rule-based NLP or even hybrid NLP, there is a strong belief that Linguistic Knowledge Bases (LKB) is the stepping stone of all.

Digital Linguistic Resources (LR) is now pursued in multiple laboratories in various standards, volumes and for a large number of languages. The success of WordNet led to a revolution in the creation of LRs with the advent of other WordNet like-type LRs for most of the European languages and with the inclusion of richer linguistic information.

An important aspect in creating LKBs is the model used and the granularity level of the linguistic information. We find LKBs that describe several aspects of a language or of language pairs using shallow or deep structures, the case of WordNet, EuroWordNet,

BalkaNet, (Lenci, 2000), (Calzolari, 2001), (Barbu, 2008). Also, we find initiatives that deal with all aspects of a language and provide deep description of linguistic information, the case of (Diaconescu, 2007). Automatic creation of LKBs is preferred but, at best, ends in obtaining large corpuses using shallow parsing. Building specialized tools that linguists operate easily is required for the management of complex language descriptions (Diaconescu et. all, 2009), (Simionescu, 2012).

Lexicon networks are a particular field in the creation of LKBs which gained, in time, a great interest as a linguistic resource showing improved outcomes for various NLP tasks. As a result, if one is to construct an LKB model this usually involves at least one lexicon network with various types of relations (synonyms, antonyms, hyponyms, etc).

A great part of lexicon networks is in fact the semantic networks. Semantic networks are a particular type of lexicon networks that implies interconnecting the meanings of lexicon entries using one or more specific types of semantic relations. However, semantic relations that make up these networks have a significant disadvantage: the directed graph obtained from a single type of semantic relation is not a connected graph, but is in fact a large set of isolated connected components. To overcome this situation, the graphs are mixed together in order to obtain a large network of interconnecting meanings. Having a large network with various types of relations can be a serious challenge when mining for knowledge in NLP tasks.

Taking into consideration the above, our work focuses on describing the process of creating a particular semantic network obtained by tagging the tokens in a lexicon's glosses with their contextual meaning. The paper is organized as follows: In Section 2, we present related research and work that make the basis of this research report. In Section 3, we further describe the annotation process and its challenges. In section 4, some measures of the annotation process are provided. Section 5 concludes the current research activity.

2. Related Work

As of 2008, the WordNet incorporates the "Princeton Annotated Gloss Corpus" - a corpus of manually annotated WordNet synset definitions ("glosses"). WordNet 3.0 is the sense inventory against which the annotation was conducted. This type of annotation implies the tokenization of each gloss (words or collocations) and manually selecting a contextual meaning for each token. In other words, a gloss is linked with other glosses by a semantic relation called "sense-tagged glosses". This type of semantic network can be very dense and it links senses belonging to different morphological lexicon entries.

A great deal of interest is shown for this type of network, as it is a natural extension of a classical concept, the 'dictionary'. While a dictionary is a collection of words explaining themselves in plain text, the "sense-tagged glosses" network is a collection of senses explaining one another in a structured way.

While in the case of WordNet it was mostly a manual undertaking, others are pursuing this using automating or semi-automating approaches either for improving WordNet, the case of eXtended WordNet, (Castillo et all, 2004), (Litkowski, 2004), (Moldovan, Novischi, 2004), or for other languages (Navigli, 2009).

As mentioned above, semantically annotated glosses show promising outcomes in various NLP tasks, especially in word sense disambiguation (Banerjee, Pedersen, 2002) (Navigli, Velardi, 2005) (Mincă, Diaconescu, 2013).

3. Manual disambiguation of glosses

The consideration of glosses as a potential linguistic resource for NLP tasks started relatively late in the field, (Lesk, 1986). Glosses were first mined for the purpose of word sense disambiguation. Two words that share the same context are disambiguated by overlapping their glosses' definitions. The meanings with a maximum overlapping score would have been selected for each of the two words. Another version of the Lesk algorithm, less complex, would overlap the glosses for each word only to the context. Soon enough, this approach outlined the importance of glosses and the Lesk reasoning was extended with the need to disambiguate the glosses (as in Figure 1).

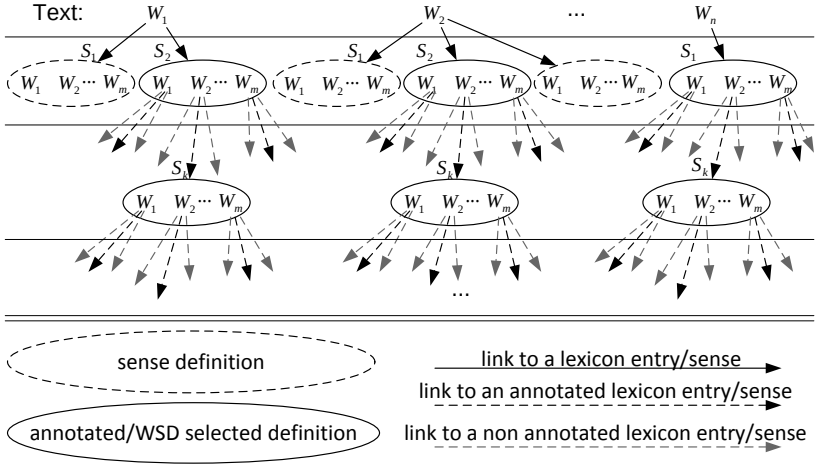

Figure 1: Lesk algorithm reasoning extended. Every annotated sense is extended with its definition that also has words with disambiguated senses and so on.

As in the case of WordNet sense-tagged glosses, we considered to manually disambiguate the glosses for Romanian language. In order to do that, a thorough study of the Romanian glosses was conducted and a specialized tool was developed for the task.

3.1. Annotation model

As seen in Figure 2, the annotation model is very similar to the one used in the case of WordNet.

However, we considered that it will be useful to not only tag the words in a gloss with their contextual meaning, but also tag each word with a contextual degree of relevance for its contribution in building up the meaning of the gloss. For this purpose, we devised three levels of relevance that a linguist can choose for a word: *Weak*, *Medium* and *Strong*. The levels of relevance have the following parity: (5, 5/4, 1).

Additional information can be added in the annotation process (lemma, morphological and syntactic categories, etc.) but it is preferable to automate the process as much as possible.

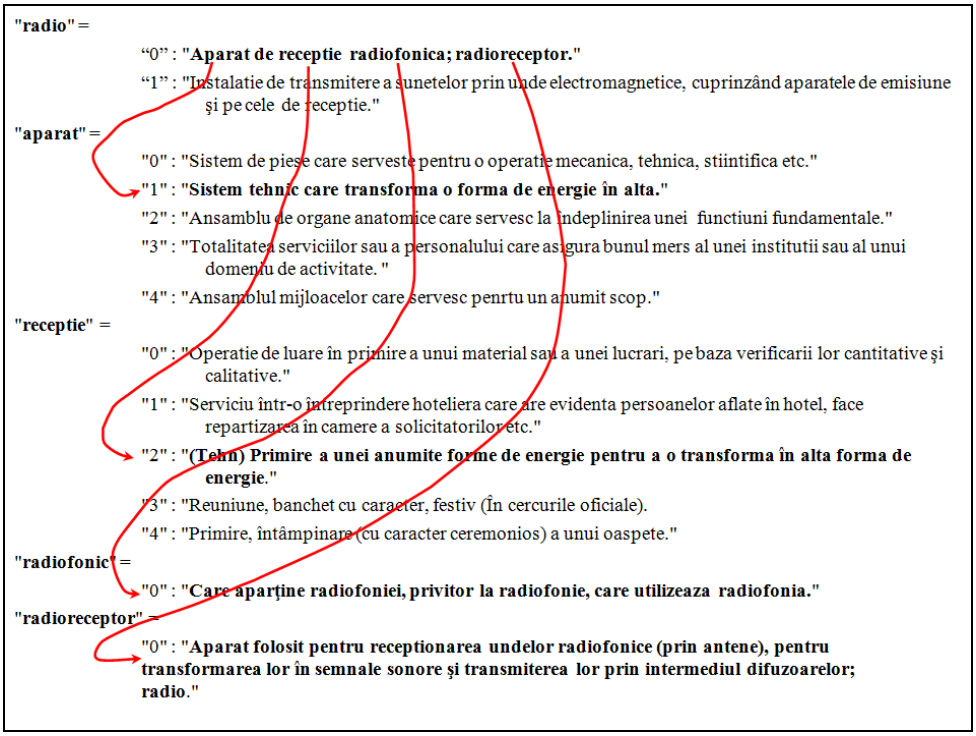

Figure 2: Annotation example for the Romanian word "radio" (first meaning)

In the case of large lexicons (like Romanian) the estimated effort is at a maximum of 300,000 glosses and a maximum of 3,000,000 tokens that need to be operated.

3.2. Sense Inventory

Our lexicon is based on the Explanatory Dictionary of the Romanian Language, 1998 (DEX - Dicţionarul explicativ al limbii române). Mostly, the glosses were imported from the digital resource "DEX online" (found at http://dexonline.ro). To fill certain gaps (missing words, missing definitions) other sources were used, but in a limited degree, like the Small Academy Dictionary (Micul Dicţionar al Academiei – 4 volumes), and the Thesaurus Dictionary of the Romanian Language (Dicţionarul Tezaur al Limbii Române – 19 volumes). Moreover, the text of the definitions was updated according to the morphological, orthoepic and ortographical rules form DOOM2. The structure of a dictionary, DEX '98 implicitly, does not fully matched the needs for our LKB model. Therefore, we used other criteria in structuring the lexicon, in deciding what entry words the lexicon should have, and, implicitly, how to divide the definitions.

Generally, dictionaries group their definitions for entry words taking into account the etymology of the word (the etymology is their main criterion in deciding whether some words with the same form are homonyms or polysemantic words), regardless of the morphological category or aspects like animation, defective of plural/singular form, gender, transitivity, reflexivity, etc. Therefore, we may find, gathered under a single

entry word, definitions for different morphological categories that word may have, as for example:

> ***FRUMÓS, -OÁSĂ,*** *frumoși, -oase,* ***adj., adv., s. n***. ***I.*** *Adj.* ***1.*** *(Adesea substantivat; despre ființe și părți ale lor, despre lucruri din natură, obiecte, opere de artă etc.) Care place pentru armonia liniilor, mișcărilor, culorilor etc.; care are valoare estetică; estetic.* ◊ *Arte frumoase = pictură, sculptură, gravură (în trecut și arhitectură, poezie, muzică, dans)...*

In contrast, our framework, GRAALAN (Diaconescu, Dumitrascu, 2007), takes into consideration the aspects mentioned above. Our lexicon has for each morphological category (considering also gender, transitivity, reflexivity, etc) an entry word in the dictionary, thus dividing the definitions from DEX '98.

In order to build a coherent and uniform semantic network, some steps had to be followed, some preceding the actual annotation process. A first step was to evaluate and to improve the definitions from the lexicon in order to comply with the annotation method and to extract some rules of good practices (see 3.3) for the annotation process. Overall, definitions suffered changes like additions, deletions, separations, updates, etc.

3.3. Good practices

Each word from a definition has a certain semantic degree of relevance depending on its contribution in defining the respective word. As it was described before, three degrees of relevance were used: ***weak, medium*** and ***strong***. There is also the posibility to ***Ignore*** various words from a definition if they do not bring any semantic contribution in defining a word. It is very important to asign the degrees of relevance in a uniform way, complying with some general rules in order to obtain a coherent semantic network.

Further on we will present the main rules observed in the annotation process, however, without going into details. In addition to the rules described below, there are also some particular rules that apply to a limited number of definitions, or to patterns of definitions that we will not mention here.

3.3.1. Words ignored in the process of annotation

As a general rule, the functional elements from a definition like certain conjunctions and prepositions, certain relative or indefinite pronouns (*care*, *cine*, *cineva*, *vreunul*) with no semantic relevance, certain adverbs (*unde*, *când*) and pronominal determiners when they do not have anaphorical value (*acest*, *acesta*), indefinite and genitive articles (*un*, *o*, *a*, *al*), the auxiliary and copulative verb *a fi* (when it has a positive form) are ignored in the process of annotation.

3.3.2. Degree of relevance

Generally, within a definition, the highest degree of relevance (***Strong***) is given to the word that has the biggest contribution in building the semantic of the word defined. Most of the times, this word is of the same morphological category as the word defined (a synonym, a word from the semantic/lexical paradigm of the word, etc.).

For example, in *CASĂ – **Clădire** care servește drept locuință,* and *PREOȚIE - Calitatea, demnitatea, funcţia de **preot,*** the words *clădire* and *preot* will have the degree of relevance ***Strong.***

The degree of relevance ***Medium*** is assigned mainly to hyperonyms or to generic words when we have abstract or verbal nouns. For example, in *BUNĂTATE - **Însușirea** de a fi bun, **înclinarea** de a face bine; PLECARE - **Acțiunea** de a pleca,* the words *însușirea, înclinarea* and *acțiunea* will be assigned ***Medium***.

The words that are asigned ***Weak*** are those with a minimum impact on the overal meaning of the word defined. Such words are, for example, from the metalanguage of the definition as *reprezintă, marchează, exprimă, aparţine,* etc, or the copulative verb *a fi* when has a negative form.

3.3.3. Identified problems during the annotation process and solutions

Two of the main problems encountered in the annotation process were the missing words or idioms from the lexicon (negative participles, composed words with prefixoids, neologisms, etc.) and the incomplete meanings. These words/idioms have been introduced in the lexicon and the definitions were enriched as they were found.

Another problem was that of the anaphoric words (a word that refers one or more previous words) as in: *TÂNĂR - Care a fost plantat sau a răsărit de puțină vreme, care n-a ajuns încă la maturitate; care este format din **asemenea** plante* where *asemenea* refers to the type of plants described before. In this case, the rule adopted was to introduce the definition *anaphoric value* to each word that may be anaphoric.

3.4. Annotation Tool

In order to facilitate the manual annotation process a dedicated tool was devised considering the following principles: to be a collaborative tool, to ensure an ergonomic but still a fast tagging task, to allow the browsing of the current network, to provide a quick feedback in the case of any changes in the network, to allow different sense mappings for a certain word in a gloss than those given by the annotator.

3.4.1. Workflow

A linguist will start working with this tool by first choosing a lexicon entry of interest. From the list of meanings corresponding to this entry, a certain meaning is chosen to have its gloss tagged. The selected gloss is tokenized and the resulting tokens are annotated with a list of meanings. In case a token is not recognized by the annotator, the linguist can provide another word form and thus recall the annotator. Then, the linguist selects groups of tokens for which will provide a certain degree of relevance or can ignore a specific group. For each valid token, the linguist will choose the contextual meaning. Finally, the linguist will save the modifications and the tool will mark the current annotated gloss as in progress or completed.

4. Statistics

The following tables show the annotation results in contrast with the measurements found in WordNet. A greater tagging density can be observed in the case of Romanian glosses.

Table 1: Romanian glosses vs. WordNet

LexNets	Glosses	Tagged Glosses	Targeted Glosses	Tags Density
Romanian	130,087	118,536	58,976	0.5757
WordNet	206,941	206,938	59,251	0.3486

Table 2: Annotation results and WordNet

LexNets	All Tokens	Operated Tokens	Operated Tokens Valid	Op. Tokens Related	Op. Tokens Valid & Related
Romanian	1,528,819	1,191,942	691,010	720,420	686,210
WordNet	2,394,190	2,394,190	2,394,189	834,803	834,803

The number of valid and related operated tokens is distributed as follows: **299,098** ***Strong*** relations, **209,879** ***Medium*** relations and **177,233** ***Weak*** relations.

The annotation process can be covered by an effort of a total of 40 man-months.

5. Conclusions

We presented here the experience gathered from manually building a lexicon network over a dictionary's glosses, for Romanian language. A dedicated tool drastically reduced the manual labor of linguists. We concluded a set of specific rules for sense-tagging of Romanian glosses that can also be considered for other languages. Finally, the effort ended with a large network of over 130,000 meanings interconnected by over 600,000 "sense-tagged glosses" relations. [1]

Acknowledgements

This work was carried out in the project SenDiS, co-funded under Romanian grant no. 207/20.07.2010.

References

Lenci A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. 2000, SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13(4): 249-263.

Calzolari, N., Lenci A., Zampolli A. 2001. International Standards for Multilingual Resource Sharing: The ISLE Computational Lexicon Working Group.

[1] Please contact the authors for information on availability of the tagged glosses and the annotation tool.

Proceedings of the Workshop on Sharing Tools and Resources, 39th ACL, 7 July 2001, Toulouse, France: 71-78.

Barbu, A.-M. 2008. Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries, *The 6th edition of the Language Resources and Evaluation Conference*

Diaconescu, St., Dumitrascu, I. 2007. Complex Natural Language Processing System Architecture, Advances in Spoken Language Technology, *The Publishing House of the Romanian Academy*, Bucharest, pp. 228-240

Diaconescu, St., Ingineru, C., Codirlasu, F., Rizea, M., Bulibasa, O. 2009. General System for Normal and Phonetic Inflection - *SpeD 2009 Conference on Speech Technology and Human-Computer Dialogue*, 18-21 June, Constanta

Simionescu, R. 2012. Graphical Grammar Studio as a Constraint Grammar Solution for Part of Speech, *ConsILR2012*

"Princeton Annotated Gloss Corpus". http://wordnet.princeton.edu/glosstag.shtml

Castillo, M., Real, F., Asterias, J., Rigau, G. 2004. The talp systems for disambiguating wordnet glosses. *In Proceedings of ACL 2004 SENSEVAL-3 Workshop*, pages 93–96, Barcelona, Spain.

Litkowski, K. C. 2004. Senseval-3 task: Word-sense disambiguation of wordnet glosses. *In Proceedings of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain: 13–16.

Moldovan, D., Novischi, A. 2004. Word sense disambiguation of wordnet glosses. *Computer Speech & Language*, 18:301–317.

Navigli, R. 2009. Using Cycles and Quasi-Cycles to Disambiguate Dictionary Glosses. *Proc. of 12th Conference of the European Association for Computational Linguistics (EACL 2009)*, Athens, Greece, pp. 594-602.

Banerjee, S., and Pedersen, T. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

Navigli, R., Velardi, P. 2005. Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 27(7)*, July, pp. 1063-1074.

Mincă, A., Diaconescu, St. 2013. An Approach to Reduce Part of Speech Ambiguity Using Semantically Annotated Lexicon Definitions. *MSIE 2013*, September 28-29. Jinan, Shandong, China.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, *Proceedings of SIGDOC '86*.

DEX – Dicţionar explicativ al limbii române, ed. a 2-a, coord. Ion Coteanu, Luiza şi Mircea Seche, Ed. Univers Enciclopedic, Bucureşti, 1998.