A Romanian SemCor Aligned to the English and Italian MultiSemCor

M. Lupu[1], D. Trandabăţ[2], M. Husarciuc[1]
[1] Faculty of Computer Science, „Al. I. Cuza" University of Iaşi, Romania
[2] Institute for Computer Science, Romanian Academy, Iaşi, Romania
E-mail: [1]mlupu@infoiasi.ro, [2]dtrandabat@iit.tuiasi.ro, [1]mhusarciuc@infoiasi.ro

Abstract. The SemCor project involved the building of a large corpus in which the words were morphologically and semantically disambiguated with senses from WordNet (1.6 and then 2.0). This achievement in the natural language processing domain (NLP) initiated another called MULTISemCor-IRST developed in Italy, which consisted in the translation of 116 texts from the SemCor corpus (a subset of the Brown Corpus). These texts were automatically word aligned and the semantic annotations were automatically transferred from the English words to their Italian translation equivalents. The next step was to extend this project enriched with Romanian translations for the 116 English texts in MultiSemCor and to create a Romanian SemCor aligned to the English and Italian one. This new project bears the name of MultiSemCor+ and it is conceived to be the test bed for multilingual semantic disambiguation experiments.

1. Introduction

As we all know, NLP domain has as a purpose the creation of programs in order to analyse, to understand and to generate languages that people use naturally, so as in the end be possible to communicate with the computer as if it were a person. Parallel texts alignment techniques are more and more necessary in domains such as translation, lexicography or terminology, where they contribute to the construction of multilingual lexical resources. To benefit from the resources available by way of unknown languages (and to learn them would take too much time), researches were made with the purpose of providing the necessary data for communicating in different languages. The MultiSemCor – IRST (*Instituto per la Ricerca Scientifica e Tecnologica-Trento*), derived from the English SemCor – achievement of Princeton University, was designed as a first step to fulfil this task.

Given the developments of statistical methods for machine learning in NLP, with controlled training methods leading to major improvements in terms of performance on various tasks, a valuable resource is now represented by large linguistically annotated corpora. Taking into consideration that up to not so long ago linguistically annotated corpora were the product of manual annotation or the manual assay of automatic annotations, it is self-evident that any attempt to reduce human effort necessary to the similar quality production of annotated data is of a great benefit to the domain. Recent studies show that a valuable opportunity could be parallel corpora which can be exploited in the creation of resources for new languages by way of annotations available in another language. This method implies the creation of a semantically annotated corpus through the exploitation of the information contained in an already annotated corpus, using the word alignment.

2. The SemCor corpus

The English SemCor, developed at Princeton University, is a subset of the English Brown Corpus and contains about 700000 words. In SemCor all the words are POS tagged and more than 200000 content words are also lemmatized and semantically annotated with reference to the WordNet lexical database (Fellbaum, 1998). The Semcor corpus is made up of 352 texts. In 186 texts all open-class words (such as nouns, verbs, adjectives and adverbs) are POS tagged, lemmatized and semantically annotated. The SemCor component of all word types consists in 359.732 tokens of which 192.639 are semantically annotated, while the verbs component has got 316.814 tokens from which 41.497 verbal occurences are semantically annotated.

2.1. The Semantic concordance

The term of “semantic concordance” is used with reference to a subset of texts in which a concept occurs. Concepts are expressed in texts by words and this way, in order to build semantic concordances, one has got to bear in mind two lexical semantics aspects: polysemy (when a given word expresses different concepts) and synonimy (when different words express the same concept). To this purpose, the WordNet database (Miller, 1990)[1] proves to be a very useful tool.

The WordNet database, created by Princeton University, is a lexical database, in which nouns, verbs, adjectives and adverbs are organized in synonym sets (*synsets*) each representing one underlying lexical concept. The synsets are linked through various relationships such as hypo/hyperonymy, meronymy and antinomy. As for semantic concordances WordNet synsets can be used to expand the word search to all its synonyms. Moreover, synset relations can be exploited in order to look for related concepts, which is not valid with normal word-based techniques. The WordNet database is frequently used as a tool for WSD (*word sense disambiguation*), both automatically and manually.

In manual WSD, the WordNet database plays the role of extensive senses inventory, such as the SemCor project in which a subset of the English Brown corpus was annotated with WordNet senses for educational use. In the past few years, within the PLN community WordNet has become the reference lexicon for almost all the tasks implying WSD (Bentivogli, Forner, Pianta).

[1] apud (Tufiş, Barbu, Ion 2004)

Therefore, within SemCor, the semantic concordance is a textual corpus linked to a lexicon with semantic tags. The concordance consists of 352 files from the Brown Corpus annotated with pointers to word senses in the WordNet database. An X Windows application, Escort, is provided for searching the concordance files for occurrences and co-occurrences of semantic tags. The semantic concordance can be seen as either a corpus in which words were syntactically and semantically tagged, or a lexicon in which the sample sentences can be retrieved for many definitions

2.2. The Brown Corpus

The **Brown Corpus of Standard American English** was the first of the modern, computer readable, general corpora. It was compiled by W.N. Francis and H. Kucera, Brown University, Providence, RI. The corpus consists of one million words of American English texts printed in 1961. The texts for the corpus were sampled from 15 different text categories to make the corpus a good standard reference. Today, this corpus is considered small, and slightly dated. The corpus is, however, still used. Much of its usefulness lies in the fact that the Brown corpus lay-out has been copied by other corpus compilers. The LOB corpus (British English) and the Kolhapur Corpus (Indian English) are two examples of corpora made to match the Brown corpus. The availability of corpora which are so similar in structure is a valuable resourse for researchers interested in comparing different language varieties, for example.At the University of Freiburg, Germany, researchers are compiling new versions of the LOB and Brown corpora with texts from 1991. This will undoubtedly be a valuable resource for studies of language change in a near diachronic perspective.

The Brown corpus consists of 500 texts, each consisting of just over 2,000 words. The texts were sampled from 15 different text categories. The number of texts in each category varies (see below).

- A. PRESS: REPORTAGE (44 texts)
- B. PRESS: EDITORIAL (27 texts)
- C. PRESS: REVIEWS (17 texts)
- D. RELIGION (17 texts)
- E. SKILL AND HOBBIES (36 texts)
- F. POPULAR LORE (48 texts)
- G. BELLES-LETTRES (75 texts)
- H. MISCELLANEOUS: GOVERNMENT & HOUSE ORGANS (30 texts)
- J. LEARNED (80 texts)
- K: FICTION: GENERAL (29 texts)
- L: FICTION: MYSTERY (24 texts)
- M: FICTION: SCIENCE (6 texts)
- N: FICTION: ADVENTURE (29 texts)
- P.FICTION: ROMANCE (29 texts)
- R. HUMOR (9 texts)

The "raw" data were reformatted and syntactically tagged (using **Eric Brill's stochastic part-of-speech tagger**) before semantic tags were assigned. After semantic tagging, the files conform to the SGML-like file format described in **cxtfile**(5WN) . The tools and programs used to create the semantic concordances are not distributed. **escort**(1WN) is a window-based browser used to search the semantic concordances for instances of semantically tagged word forms. It can be used to find semantic tags to one or more senses of a word and optional co-occurring senses.

2.3. Semantic Concordance Organization

The semantically tagged Brown Corpus files are divided into three semantic concordances based on what was tagged and when. Each is stored in a separate directory by the concordance's name (*conc*). The concordances are:

conc	**Contents**	**What's Tagged**
brown1	103 Brown Corpus files	All open class words
brown2	83 Brown Corpus files	All open class words
brownv	166 Brown Corpus files	Verbs

Each file is named using the following convention: **`br-`***`article_code`* where *`article_code`* is a letter followed by a two digit number that denotes the section and article number that the text was derived from. No file is in more than one **semantic concordance.**

3. MultiSemCor

The parallel annotated corpus within the MultiSemCor corpus is created by the exploitation of the English SemCor corpus (Landes et al., 1998). MultiSemCor consists in 116 English texts along with their corresponding 166 Italian translations. The texts have been automatically aligned at word level and the English semantic annotations have been automatically transferred from the English words to their Italian translation equivalents.

The corpus consists in:

- 116 English SemCor texts annotated with sentence and paragraph splitting, morphosyntactic information, word sense information (WordNet 1.6 synsets);
- 116 Italian translations of the English SemCor texts annotated with sentence and paragraph splitting, morphosyntactic information, word sense information (automatically transferred from English, MultiWordNet Italian synsets)
- 116 word alignment files representing:
    - sentence alignment
    - word alignment

The corpus is encoded using XML as a common logic data format and taking into account, whenever possible according to the requirements of our NLP applications, the Corpus Encoding Standard guidelines and the new standard ISO/TC 37/SC 4 for language resources. For more consideration, see[2]

a) English files

Each English file contains all the information present in the original SGML SemCor file; annotations are represented with the MultiSemCor XML format, which requires nestable "structure" and "feature" elements containing "type" and "id" attributes with different values.

b) Italian files

In each Italian file, morphosyntactic and semantic annotations are represented with the MultiSemCor XML format, which requires nestable "structure" and "feature" elements containing "type" and "id" attributes with different values.

The morphosyntactic annotation is carried out using ITC-irst NLP tools working on Italian, while the semantic annotation is automatically transferred from the corresponding words in the English original SemCor text.(Bentivogli, Forner, Pianta)

c) Alignment files

Each alignment file contains the alignment at the sentence and word level of the English text and its Italian translation; annotations are represented following XCES guidelines, which require "linkGrp", "link", and "align" elements. Xlinks and Xpointers refer respectively to the monolingual English and Italian files.

5. Romanian SemCor

In building the Romanian Semcor (based on the English SemCor), several stages have been covered.

- Translation

34 English SemCor textes have been translated into Romanian consisting in 65926 tokens in 3871 sentences. The English SemCor sentence and paragraph annotations have been obseved and diacritics have been converted into SGML entities. At this stage, there had to be used the underscore character with respect to phrasal verbs, idioms, complex proper nouns and collocations as to be taken as a single lexical unit. This aspect is of most importance in retrieving the translation equivalence in parallel texts (*bitexts*).

- Preliminary processing

[2] http://www.cs.vassar.edu/CES/
http://www.cs.vassar.edu/XCES/
http://www.tc37sc4.org/

Segmentation: recognizing phrases/collocations made up of more than one word as single lexical tokens and splitting single words into multiple lexical tokens (when it is the case). The program performing this task is called a segmenter or tokenizer. Words and phrases consisting in more than one word are to be designated as lexical tokens or simply tokens. (Tufiş, Barbu, Ion) The segmenter used at this preliminary stage is designed by RACAI and it won the first prize in ACL2005 Workshop.

Sentence alignment: RACAI managed to make alignment units (translation units) tally. The procedure is to be found in (Tufiş, Barbu, Ion 2004)

Tagging and lemmatizing stage. For morphosyntactic disambiguation RACAI made use of a row tagging method (*tiered-tagging*) with combined language models (Tufiş, 1999) Tnt based – a HHH trigram tagger (Brants, 2000).

- Processing stage

A TECL (translation equivalence candidates list) has been computed at this stage, along with a filtering of noise data using the *loglikelyhood* stochastic test (LL).

- Sense importation

In the MultiSemCor+ project, word sense correspondence between Romanian SemCor and English SemCor is automatically achieved (using sense annotations). It's important to know that sense annotations in the Romanian corresponding files are the same as in SemCor and the synset id is taken from the WordNet 2.0 database. This was an obstacle in mapping the senses in MultiSemCor as it uses the WordNet 1.6 version.

- The Romanian SemCor corpus realisation

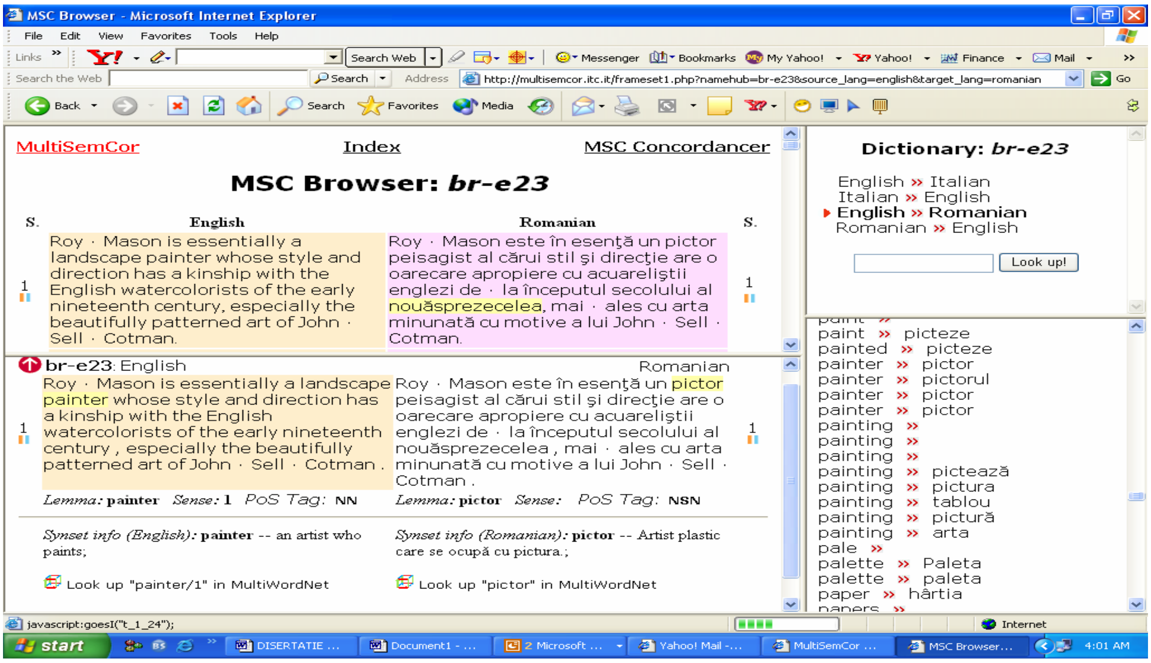
Romanian SemCor consists in:

- 34 English SemCor translations automatically annotated with sentence and paragraph splitting, morphosyntactic information, sense information (automatically transferred on MultiWordNet English synsets);
- 12 word alignment files alignment files representing sentence alignment and word alignment
- 22 files in progress of alignment

Accessing the Romanian SemCor on line

Opening the MultiSemCor web page[3] and choosing the option *Browser*at the left handside of the page, we enter a site disposed in four windows. The first one entitled **MSC Browser: index** gives one the opportunity to choose between two couples of languages: *english-italian* or *english-romanian*. On the left handside are displayed the numbers of the texts in MultiSemCor. By pointing to the *Aligned* tab there will be posted, on two rows, the English and Romanian sentence-aligned equivalent texts. On pointing

[3] http://multisemcor.itc.it/


MSC Browser - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links Search Web Messenger Bookmarks My Yahoo! Yahoo! Finance Mail
Search the Web Search Address
http://multisemcor.itc.it/frameset1.php?namehub=br-e23&source_lang=english&target_lang=romanian
Go
Back Search Favorites Media
MultiSemCor
Index
MSC Concordancer
javascript:goesI("t_1_24");
Internet
start
DISERTATIE ...
Document1 - ...
2 Microsoft ...
Yahoo! Mail -...
MultiSemCor ...
MSC Browser...
4:01 AM

MSC Browser: *br-e23*

S.	English	Romanian	S.
1	Roy · Mason is essentially a landscape painter whose style and direction has a kinship with the English watercolorists of the early nineteenth century, especially the beautifully patterned art of John · Sell · Cotman.	Roy · Mason este în esenţă un pictor peisagist al cărui stil şi direcţie are o oarecare apropiere cu acuareliştii englezi de · la începutul secolului al nouăsprezecelea, mai · ales cu arta minunată cu motive a lui John · Sell · Cotman.	1

br-e23: English Romanian

	English	Romanian	
1	Roy · Mason is essentially a landscape painter whose style and direction has a kinship with the English watercolorists of the early nineteenth century , especially the beautifully patterned art of John · Sell · Cotman .	Roy · Mason este în esenţă un pictor peisagist al cărui stil şi direcţie are o oarecare apropiere cu acuareliştii englezi de · la începutul secolului al nouăsprezecelea , mai · ales cu arta minunată cu motive a lui John · Sell · Cotman .	1

Lemma: **painter** *Sense:* **1** PoS Tag: **NN**

Lemma: **pictor** *Sense:* PoS Tag: **NSN**

Synset info (English): **painter** -- an artist who paints;

Synset info (Romanian): **pictor** -- Artist plastic care se ocupă cu pictura.;

Look up "painter/1" in MultiWordNet

Look up "pictor" in MultiWordNet

Dictionary: *br-e23*

English » Italian
Italian » English
▸ **English » Romanian**
Romanian » English

Look up!

paint	»	picteze
painted	»	picteze
painter	»	pictor
painter	»	pictorul
painter	»	pictor
painter	»	pictor
painting	»	
painting	»	
painting	»	
painting	»	pictează
painting	»	pictura
painting	»	tablou
painting	»	pictură
painting	»	arta
pale	»	
palette	»	Paleta
palette	»	paleta
paper	»	hârtia

Fig.1 MultiSemCor interface

any word/collocation in the bitext the word alignment for both Romanian and English words will be displayed in the *Sentence alignments* window below providing PoS and MultiWordNet synsets information. On the righthandside there are two windows one providing a dictionary and the other a term displaying window in which the pointed word is aligned to its translation equivalent (Fig.1).

It is intended to have the Romanian translations aligned to their Italian equivalents by using the English semantic annotations in the near future.

The internet address http://multiwordnet.itc.it/online/ opens a page dedicated to finding the synsets in English, Italian, Spanish, Hebrew and newly Romanian. http://multisemcor.itc.it/frameset2.php opens another window where, by introducing a word and its lemma, it is possible to visualize the contexts in which it appears, together with information about its PoS and synset. If the word has not been aligned yet, there will come up a message in red *This word has not been* aligned. Pointing to a word in this window, we are redirected to the **MSC Browser: index**.

5. Conclusions

This web page, the offspring of the shared efforts belonging to Romanian, Italian and American researchers from well-known universities, is rich in various word sense disambiguation information. It is useful not only for individual purposes, e.g. learning Romanian, Italian or English via one /two of the mentioned languages, but also for research purposes in the PLN related informational domain.

Parallel texts are an essential method towards the creation of new resources for new languages via prior annotations available in a different language. The English SemCor represented a turning point in the PLN investigations by offering the necessary technology to create a morpho-syntactically and semantically annotated corpus, a word level and semantically aligned corpus. MultiSemCor, by way of sense mapping from English to Italian freed the way towards more laborious researches whose ultimate aim is WSD. The Romanian SemCor comes up from the wish to take an active part in the PLN investigations, making a step forward by experimenting new techniques with the principal objective as to determine computers to understand how natural language comprehension functions in human communication.

Moreover, MULTISemCor+ will constitute the major basis in the AI and NLP tests where there will be made important steps in domains such as *machine translation, machine learning, information extraction*, *word sense disambiguation – WSD*, *question answering* and finalyy but not least *speech processing.*

References

[1]. Luisa Bentivogli, Pamela Forner and Emanuele Pianta. "Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus". In Proceedings of COLING 2004, Geneva, Switzerland, August 23-27, 2004, pp. 364-370.

[2]. Emanuele Pianta, Luisa Bentivogli. "Knowledge Intensive Word Alignment with KNOWA". In Proceedings of COLING 2004, Geneva, Switzerland, August 23-27, 2004, pp. 1086-1092.

[3]. Marcello Ranieri, Emanuele Pianta and Luisa Bentivogli. "Browsing Multilingual Information with the MultiSemCor Web Interface". In Proceedings of the LREC 2004 Workshop "The Amazing Utility of Parallel and Comparable Corpora", Lisbon, Portugal, May 2004, pp. 38-41.

[4]. Emanuele Pianta and Luisa Bentivogli. "Translation as Annotation". In Proceedings of the AI*IA 2003 Workshop "Topics and Perspectives of Natural Language Processing in Italy", Pisa, Italy, September 2003, pp. 40-48.

[5]. Emanuele Pianta, Luisa Bentivogli and Christian Girardi."MultiWordNet: developing an aligned multilingual database". In Proceedings of the First International Conference on Global WordNet, Mysore, India, January 21-25, 2002.

[6]. Christiane Fellbaum (ed). WordNet: An Electronic Lexical Database. Cambridge(Mass.), The MIT Press, 1998.

[7]. D. Tufiş, D. Cristea, S. Stamou. BalkaNet: “Aims, Methods, Results and Perspectives. A General Overview”. Romanian Journal of Information and Technology, Volume 7, Numbers 1-2, 2004, 9-43

[8]. D. Tufiş, Ana Maria Barbu and Radu Ion. “Extracting multilingual lexicons from parallel corpora”. *Computers and the Humanities Volume 38, Issue 2, May 2004, pp. 163-168©2004. kluwer Academic Publishers. Netherlands*

[9]. Miller, G.A., Leacock, C., Tengi, R., and Bunker R. T. (1993). A Semantic Concordance, *"Proceedings of the ARPA WorkShop on Human Language CEThnology"* . San Francisco, Morgan Kaufman.