Ontology-Supported Text Classification Based on Cross-Lingual Word Sense Disambiguation

Dan Tufiş[1] and Svetla Koeva[2]

[1] Research Institute for Artificial Intelligence, Romanian Academy,
13, "13 Septembrie", 050711, Bucharest, Romania
[2] Institute for Bulgarian Language, Bulgarian Academy of Sciences,
52 Shipchenski prohod, 1113 Sofia, Bulgaria
tufis@racai.ro, svetla@mail.ibl.bas.bg

Abstract. The paper reports on recent experiments in cross-lingual document processing (with a case study for Bulgarian-English-Romanian language pairs) and brings evidence on the benefits of using linguistic ontologies for achieving, with a high level of accuracy, difficult tasks in NLP such as word alignment, word sense disambiguation, document classification, cross-language information retrieval, etc. We provide brief descriptions of the parallel corpus we used, the multilingual lexical ontology which supports our research, the word alignment and word sense disambiguation systems we developed and a preliminary report on an ongoing development of a system for cross-lingual text-classification which takes advantage of these multilingual technologies. Unlike the keyword-based methods in document processing, the concept-based methods are supposed to better exploit the semantic information contained in a particular document and thus to provide more accurate results.

Keywords: cross-lingual document classification, multilingual lexical ontology, parallel corpora, word alignment, word sense disambiguation.

1 Introduction

The recent advancements in corpus linguistics technologies, as well the availability of more and more textual data, demonstrated that various well established monolingual applications could achieve a higher level of accuracy when performed on parallel data. This is not surprising as human translators incorporate a great deal of linguistic and world knowledge into their translations and when this knowledge is (even partially) revealed, it represents an exceptionally useful resource for better solving challenging NLP tasks. For instance, word sense disambiguation (WSD), a very difficult task (AI-complete), has been shown to achieve superior accuracy when done on a parallel document than in a monolingual text.

This paper is organized as follows: in section 2 we introduce the parallel corpus we work with, part of a 22-languages parallel corpus, and we shortly describe the lexical ontology we rely on in processing parallel corpora. In section 3 we give

F. Masulli, S. Mitra, and G. Pasi (Eds.): WILF 2007, LNAI 4578, pp. 447–455, 2007.

an overview of the word aligning and word sense disambiguation procedures which are highly instrumental to many NLP hard problems. In section 4 we will report on an ongoing research on a concept-based document classification system. Finally, we draw some conclusions and outline future work plans.

2 JRC-Acquis and the Aligned Wordnets

Parallel corpora became recently one of the most required language resources, because they have been proved to be essential for the development of several multilingual applications such as statistical machine translation (including translation consistency checking), multilingual categorisation, extraction of multilingual dictionaries, aligning lexical ontologies, training and testing of the multilingual information extraction software and many others. JRC-Acquis [1] is a unique parallel corpus as far as the number of languages contained (21 languages) and size of the monolingual texts (an average of more than 9 million words per languages).

An additional feature of the JRC-Acquis is the fact that most texts have been manually classified into subject domains according to the EUROVOC thesaurus, which is a classification system with over 6000 hierarchically organised classes. The JRC-Acquis parallel corpus was sentence aligned for all the language pairs (210) and it is a public resource (http://wt.jrc.it/lt/acquis/), already at the version 2.2. Although the number of documents in individual languages is almost 20,000, the JRC-Acquis distribution contains a subset of the Acquis Communautaire documents because not all the existing documents are translated in all the languages. We recently created a trilingual corpus (Bg-En-Ro) containing 16291 files in Bulgarian, 7972 in English and 18291 in Romanian. The set of English documents was extracted from JRC-Acquis version 2, while the documents for Bulgarian and Romanian were downloaded from the CCVISTA server of the Technical Assistance Information Exchange Office in Brussels. The number of documents available in all three languages was 7420. We extracted various statistics from each file for all three languages and we eliminated the documents the statics of which did not correlate in the three languages. We took into account the number of paragraphs and words. The number disparities occurred because in the JRC-Acquis the annexes were eliminated while on the CCVISTA server, the documents are complete. So, we automatically filtered out the Ro and Bg documents that, unlike the En documents, included the annexes. The final number of retained documents was 4880. Table 1 displays quantitative information for the trilingual parallel corpus, before and after the correlation filtering.

Table 1. Bg-En-Ro parallel corpus before (B) and after (A) filtering

Language	Bg(B)	En(B)	Ro(B)	Bg(A)	En(A)	Ro(A)
docs	7420	7420	7420	4880	4880	4880
pars	775504	446020	820569	357654	299486	348168
words	9747796	88821220	9844904	5849462	6046003	5784323

The Romanian and Bulgarian documents were available in MS Word format and we converted them in the xml TEI format of the JRC distribution (Fig. 1). The parallel documents of our sub-corpus were sentence aligned, tokenized, lem-

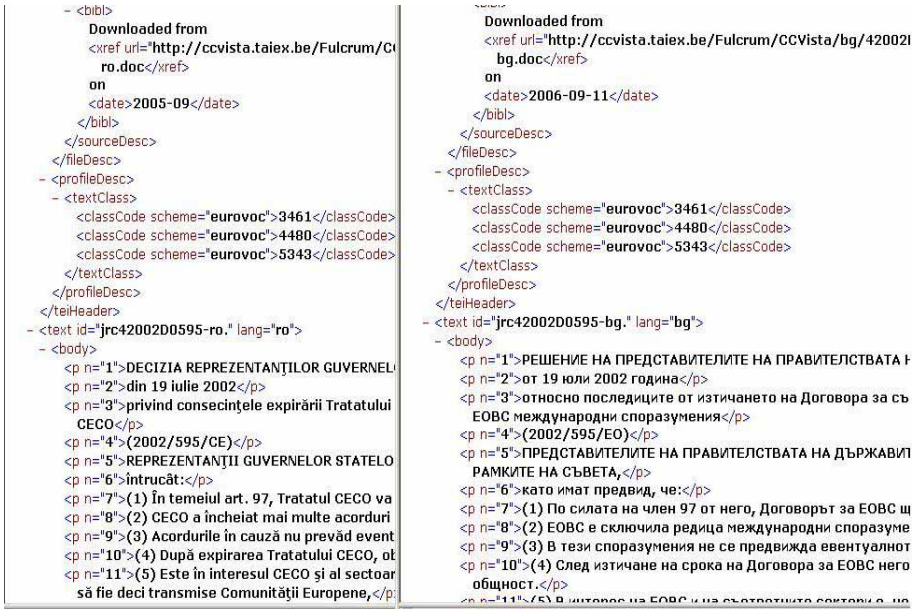

Fig. 1. Documents (Ro, Bg) encoded in compliance with the JRC-Acquis format

matized, and tagged. The tagset used is MULTEXT-EAST (nl.ijs.si/ME/) compliant. The multilingual XML encoding, exemplified in Figure 2, was inspired by XCES-Ana-Align specifications (http://www.xml-ces.org/). The XCES-Ana-Align format is the standard input for our word alignment and word disambiguation platform which will be briefly described in the section 3.

The BALKANET European project [2] created a collection of interlingually aligned wordnets for Bulgarian, Czech, Greek, Romanian, Serbian and Turkish languages, following the basic principles of EUROWORDNET [3]. The InterLingual Index (ILI) of the BALKANET wordnets is the Princeton WordNet2.0 (PWN2.0) [4]. Due to the projection of the Suggested Upper Merged (SUMO) Ontology [5] over PWN2.0, and by the multilingual equivalence linking of the BalkaNet monolingual wordnets to the PWN2.0, the SUMO/MILO labelling and inference rules are directly available to any synset of any monolingual wordnet.For instance the PWN2.0 synset (*permit*:1, *allow*:2, *let*:3, *countenance*:1) tagged by the SUMO/MILO category confersRight[1] has the ID ENG20-00776433-v which uniquely identifies the Bulgarian synset (*pozvolyavam*:3,

[1] This is a base ontology relation described in SUMO as: "%2 %n {doesn't} &%allow %p{s}%3 to perform task of the type %1".

razreshavam:3) and the Romanian synset (*încuviinţa*:1, *îngădui*:1.1.1, *permite*: 1.2). Although both Bulgarian and Romanian wordnets are significantly smaller (currently they have about 30,000 and respectively 40,000 synsets) than Princeton WORDNET (more than 115,000 synsets) due to the development strategies adopted by the BALKANET consortium, the general texts coverage is very high, as most usual words are encoded in our wordnets.

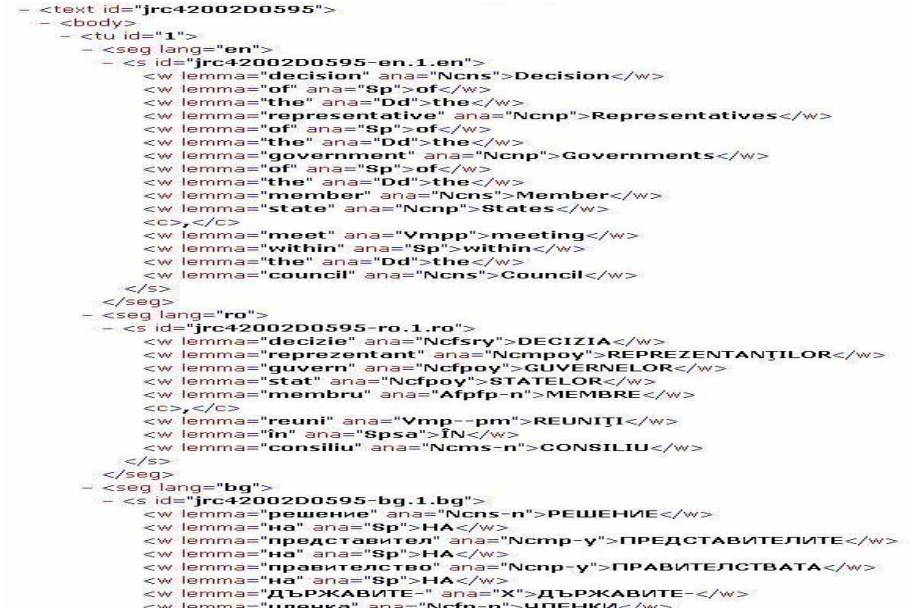

Fig. 2. The sentence aligned, tagged and lemmatized trilingual (En-Ro-Bg) corpus

3 Word Alignment and Word Sense Disambiguation

We developed an automatic procedure for word sense disambiguation in parallel texts that takes advantage of the way the words in one language were translated in the other languages. Revealing the translators knowledge embedded in the parallel texts is achieved by a highly accurate statistics-based sentence and word alignment system[2], described elsewhere [6].

The word alignment system uses a statistical alignment model and a statistical translation dictionary. For the statistical translation dictionary we use GIZA++

[2] The alignment system is called COWAL, and was the best rated in the ACL 2005 Romanian-English shared task on word alignment (see: Martin, J., Mihalcea, R., Pedersen, T.: Word Alignment for Languages with Scarce Resources. In Proc. of the ACL Workshop on Building and Exploiting Parallel Texts: Data Driven Machine Translation and Beyond, Ann Arbor, MI (2005), Figure 2).

(freely available at http://www.fjoch.com/GIZA++.html) and lemmatized parallel corpora (due to strong inflectional character of Romanian and Bulgarian, in order to increase statistical confidence, the translation equivalence probabilities are computed for lemmas not for wordforms). The alignment model consists of various weights and thresholds for different features and they are supposed to work for most Indo-European languages (cognates, translation equivalence entropy, POS-affinities, locality etc.). Based on our previous translation Ro-En model and the Ro-En translation dictionary extracted from the JRC-Acquis subcorpus described in section 2, we aligned several Ro-En parallel documents. From the Bg-En sub-corpus we extracted a Bg-En translation dictionary but since we do not have yet a Bg-En word-alignment model we used the model built for English-Romanian alignment (see `http://www.cse.unt.edu/~rada/wpt05/`). Using the same alignment model (which was tuned for Ro-En parallel texts) for Bg-En parallel texts was motivated partly because alignment model tuning for a new pair of languages is a highly demanding task and partly because, distributionally, Ro and Bg are quite similar, in spite of belonging to different language families. Obviously, in this case, the word alignment accuracy for Bg-En parallel texts is lower than the alignment accuracy for the Ro-En parallel texts but not significantly lower (preliminary estimations at the time of this writing show an F-measure around 65%). The Bg-En texts, word-aligned this way, will be partially hand-validated and corrected (where necessary) by means of a specialized editor, part of the word-alignment and WSD platform (see Figure 3). With the corrected Bg-En lexical alignment used as training data, the alignment model will be finer tuned, with direct consequences in increased alignment accuracy of the entire corpus. We generated the Ro-Bg word alignment using the transitivity

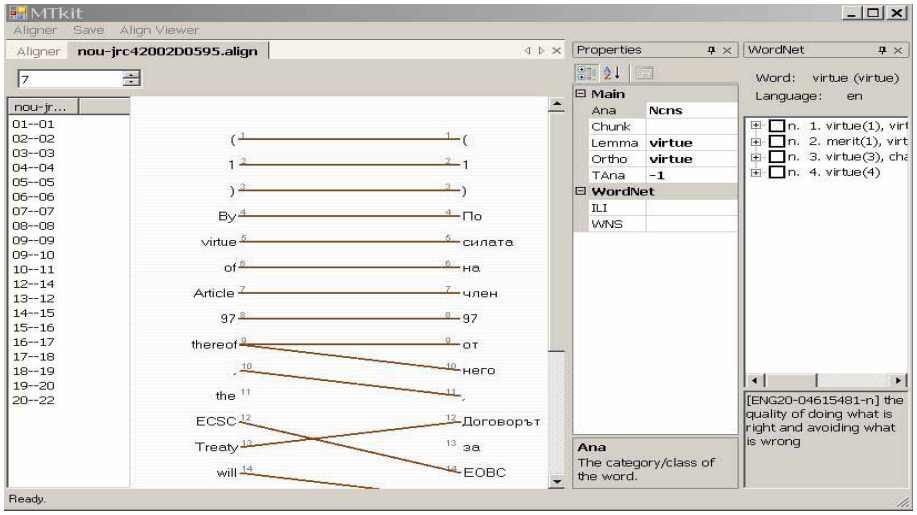

Fig. 3. The Bulgarian-English word-alignment

of the alignment links from Ro-En and En-Bg to derive the Ro-Bg alignment. The word alignment links are representations of translational equivalence between the respective tokens and we rely on the heuristics according to which if M words in language L1 are aligned to N words in the hub language L2, and these N words are aligned to Q words in language L3, then it is highly probable that the N words in language L1 are aligned to the Q words in language L3. We decided to take the hub language approach instead of the direct Bg-Ro approach for multiple reasons: it is simpler to extend to all the languages in the JRC-Acquis; for evaluations and corrections is easier to find experts understanding English and the other language; linguistic resources and the processing tools available for English, as well as the ever improving alignment technologies allow for cross-lingual annotation transfer and thus rapid prototyping of linguistic knowledge for the target language, etc. Once the parallel texts are word-aligned, the word sense disambiguation for the aligned words becomes straightforward when a multilingual lexical ontology is available for the concerned languages. Given a translation equivalence pair found by the word aligner, such as (*pozvolyavam îngădui*) the WSD system looks for unique identifiers common to the synsets containing the words *pozvolyavam* and *îngădui* respectively. If such a unique identifier is found the problem is solved: the common word sense is given by that unique ID (e.g. ENG20-00776433-v) or by the SUMO/MILO category of the unique ID (e.g. confersRight) depending on the required WSD sense granularity. It is obvious that the coarser the sense granularity, the higher the accuracy of WSD is. When using the SUMO/MILO sense inventory our WSD system has an average F-measure of more than 80%. However, as one would expect, most errors occur for the words with fewer occurrences in the corpus or for the words with a large number of distinct senses. In order to reduce the influence of semantic tagging errors we decided to consider for further processing only those words the senses of which occurred a minimal number of times (the threshold is an empirical value, depending on the documents size and the number of classes to be used in the document categorization). We also disregard words with too many senses (irrespective of their frequencies). If one agrees on the hypothesis that a word with a large number of senses is likely to be found in almost any document long enough, then it follows that the respective word would be a poor class discrimination feature for a classification system. One could see in these restrictions an analogy with the TF/IDF algorithm. With these two selectional restrictions, we estimate that the semantic tagging errors would hardly affect the final document classification performance. The reason for the optimism stems from the inherent smoothening achieved by the selection procedure of the most discriminating concepts. Our method is likely to be effective if the processed documents are not very short. For short documents one cannot afford filtering out too many words and cannot expect to see many repeated concepts. In [7] there are described the major difficulties in abstracts clustering and the authors present an interesting method to overcome these difficulties. We believe that their approach, which relies on monolingual data, could be nicely extended with a WSD method as presented here, for processing multilingual abstracts.

4 Concept-Based Text Classification

Unlike the keyword-based methods in document processing, the concept-based methods are supposed to better exploit the semantic information contained in a particular document and thus to provide more accurate results.

Having a set of before-hand classified documents, they are word sense disambiguated in terms of SUMO/MILO, and a number of concepts are selected as described in the previous section. Then, we measure the discriminative power of the selected concepts with respect to the thematic categorization of documents in terms of the majority logic operators more than n, at least n [8][3] etc. The normalized values for these thresholds give the minimum terms density of specialized lexis in the thematic reference corpora. Once these values are established, new parallel documents can be classified. Obviously, the minimum density of specialized lexis is dependent on the number of classes used for classification as well as on the sharpness of the domains differences.

In Table 2 there are summarized the observations from a preliminary experiment carried on Bulgarian data, manually sense annotated [9], where a given document could be accurately classified as belonging to one of four domains (Law, Politics, Economy and Medicine) if the density of concepts specific to that domain was at least 5%. Based on this Wizard-of-Oz experiment, we found strong motivations to automate the hand annotation part and implement a classification system based on the experimental findings. One should note that the classification mechanism based on SUMO/MILO sense tagging is more powerful than an alternative solution relying on semantic distances among the word senses in the underlying wordnets. This is because traversing wordnet relations (a paradigmatic approach) would consider semantic relatedness only among words of the same grammatical category (due to the wordnet structuring principles). The SUMO/MILO concepts labelling the wordnet synsets are insensitive to part of speech of the respective synsets. That is to say that the same SUMO/MILO concept may label words of different parts of speech (eg. the noun blow to the verb kick).

However, the wordnet structuring is complementing the SUMO/MILO ontology support. The new documents content words which are found paradigmatically related to the words tagged by a SUMO concept add relevance to the

Table 2. Frequency of domain specific lexis

Corpus vs. Lexis	Law	Politics	Economy	Medicine	Law+Economy	Law+Medicine
Law	10.3	5.4	4.2	1.1	8.4	5.9
Politics	2.9	8.3	3.4	0.5	2.9	3.6
Economy	1.2	2.0	9.2	0.6	6.5	2.5
Medicine	0.1	0.1	0.1	6.7	0.1	5.2

[3] The thresholds prescribed by these operators leave out most of the infrequent words liable to wrong disambiguation.

respective category. To this end, symmetric relations (antonymy), symmetric and transitive relations (also see, verb group, similar to), reverse relations (cause, derived, derivative, participle), and hierarchical relations (hyponymy, holonymy, subevent) are used to increase the density of the domain discriminating concepts. Thus for a given SUMO/MILO concept a basic list of referents is built from all the words occurrences tagged with the same concept. This list is extended with the words which are found to be paradigmatically related to the words already in the basic list. The number of words in this extended list, divided by the total number of the content words in the document to be classified represents the density of the respective concept. The category of a document is determined by the densities of the concepts in each lexis.

5 Conclusions and Further Work

The concept-based bitexts clustering and/or classification is a very promising application area of parallel data exploitation. One of the greatest advantages of our approach is that it can be used to automatically classify documents in several languages at once. That is, if we have a parallel corpus in multiple languages (such as JRC-Acquis corpus), word sense disambiguation and classification performed on any pair of them propagates to the rest via documents translation equivalence and the word alignment linkages.

We plan to evaluate the effectiveness of our method on the Bg-Ro subcorpus of the JRC-Acquis+. The Bg-Ro bitexts of JRC-Acquis+ will be automatically classified with the described method and the results will be compared to the present CELEX-based classification, used as reference data. The evaluation of the results will allow us to quantitatively evaluate the accuracy of our procedure and to detect any potential human classification errors in the CELEX database. Due to language independence of the CELEX classification, finding and correcting such classification errors will be beneficial for all the 22 languages present (now or in the future) in the JRC-Acquis+ parallel corpus.

References

1. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: proceedings of the 5th LREC Conference, Genoa, pp. 2142–2147 (2006)
2. Tufiş, D. (ed.): Special Issue on the BalkaNet Project. Romanian Journal of Information Science and Technology, vol. 7(1-2), Bucharest (2004), `http://www.racai.ro/BalkanetSpecialIssue.doc`
3. Vossen, P. (ed.): A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1998)
4. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
5. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In: Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine (2001)

6. Tufiş, D., Ion, R., Ceauşu, Al., Ştefănescu, D.: Improved Lexical Alignment by Combining Multiple Reified Alignments. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, pp. 153-160 (2006)
7. Alexandrov, M., Gelbukh, A., Rosso, P.: An Approach for Clustering Abstracts. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 275–285. Springer, Heidelberg (2005)
8. Pacuit, E., Salame, S.: Majority logic. In: KR Proceedings, pp. 1–26 (2004)
9. Stoyanova, I., Koeva, S., Lesseva, S.: Applying and analysing Brown corpus model for Bulgarian (to appear)