The Szeged Corpus: a POS tagged and syntactically annotated Hungarian natural language corpus

Dóra CSENDES
University of Szeged Department of Informatics
Árpád tér 2.
Szeged, Hungary, H-6720
dcsendes@inf.u-szeged.hu

János CSIRIK
University of Szeged Department of Informatics
Árpád tér 2.
Szeged, Hungary, H-6720
csirik@inf.u-szeged.hu

Tibor GYIMÓTHY
University of Szeged Department of Informatics
Árpád tér 2.
Szeged, Hungary, H-6720
gyimothy@inf.u-szeged.hu

Abstract

The Szeged Corpus is a manually annotated natural language corpus currently comprising 1.2 million word entries, 145 thousand different word forms, and an additional 225 thousand punctuation marks. With this, it is the largest manually processed Hungarian textual database that serves as a reference material for research in natural language processing as well as a learning database for machine learning algorithms and other software applications. Language processing of the corpus texts so far included morpho-syntactic analysis, POS tagging and shallow syntactic parsing. Semantic information was also added to a pre-selected section of the corpus to support automated information extraction.

The present state of the Szeged Corpus (Alexin et al., 2003) is the result of three national projects and the cooperation of the University of Szeged, Department of Informatics, MorphoLogic Ltd. Budapest, and the Research Institute for Linguistics at the Hungarian Academy of Sciences. Corpus texts have gone through different phases of natural language processing (NLP) and analysis. Extensive and accurate manual annotation of the texts, incorporating over 124 person-months of manual work, is a great value of the corpus.

1 Texts of the Szeged Corpus

When selecting texts for the Szeged Corpus, the main criteria was that they should be thematically representative of different text types. The first version of the corpus, therefore, contains texts from five genres, roughly 200 thousand words each. Due to its relative variability, it serves as a good reference material for natural language research applications, and proves to be large enough to guarantee the robustness of machine learning methods. Genres of Szeged Corpus 1.0 include:

- fiction (two Hungarian novels and the Hungarian translation of Orwell's *1984*)
- compositions of 14-16-year-old students
- newspaper articles (excerpts from three daily and one weekly paper)
- computer-related texts (excerpts from a *Windows 2000*[1] manual book and some issues of the *ComputerWorld, Számítástechnika* magazine)
- law (excerpts from legal texts on economic enterprises and authors' rights).

Text type	Size			Ratio of ambig. words
	Number of words	Number of punct. marks	Number of ambig. words	
Fiction	187191	47968	100968	53.94%
Composi-tions	223058	47208	127821	57.30%
Newspaper	187334	32948	91375	48.78%
Computer	181980	31906	82698	45.44%
Law	222186	38715	106581	47.97%
Short business news	188345	25817	82813	43,7%
Total	**1190094**	**224562**	**592256**	**49.52%**

Table 1. Data about Szeged Corpus 2.0

During further developments, the first version of the corpus was extended with a 200 thousand-word-long sample of short business news[2]. The

[1] Kis, Balázs: *Windows 2000*, Szak kiadó, 2000

[2] Short business news originate from the archive of the Hungarian News Agency (http://www.mti.hu/).

newly added section served as an experimental database for learning semantic frame mapping to be later integrated in an IE technology. Table 1. shows data referring to Szeged Corpus 2.0.

2 Annotation of the Szeged Corpus

Morpho-syntactic analysis and POS tagging of the corpus texts included two steps. Initially, words were morpho-syntactically analysed with the help of the Humor[3] automatic pre-processor. The program determined the possible morpho-syntactic labels of the lexicon entries, thereby creating the ambiguous version of the corpus. After the pre-processing, the entire corpus was manually disambiguated (POS tagged) by linguists. For the tagging of the Szeged Corpus, the Hungarian version of the internationally acknowledged MSD (Morpho-Syntactic Description) scheme (Erjavec, Monachini, 1997) was selected. Due to the fact that the MSD encoding scheme is extremely detailed and refined (one label can store information on up to 17 positions), there is a large number of ambiguous cases, i.e. one word is likely to have more than one possible labels. Experiences show that by applying the MSD encoding scheme, roughly every second word of the corpus is ambiguous. Disambiguation, therefore, required accurate and detailed work cumulating up to 64 person-months of manual annotation. Currently all possible labels as well as the selected ones are stored in the corpus.

A unique feature of the corpus is that parallel to POS tagging, users' rules have been defined for each ambiguous word in a pre-selected (202 600-word-long) section of the corpus. The aim of applying users' rules was to mark the relevant context (relevant set of words) that determines the selection of a certain POS tag. Users' rules apply $\texttt{before}_1$, $\texttt{before}_2$, ... $\texttt{after}_1$, $\texttt{after}_2$, ... predicates for marking the relevant context of a word[4]. The manually defined rules can then be generalised to regular disambiguation rules applicable to unknown texts as well. Out of the selected 202 600 words 114 951 were ambiguous. Annotators defined users' rules for these cases among which 26 912 different ones were found. The major advantage of the defined rules lies in their accuracy and specificity, wherefore they are an interesting and valuable source of additional linguistic information that can e.g. support the more precise training of machine learning algorithms.

After the completion of POS tagging, a project[5] was initiated to encompass shallow syntactic parsing of the Szeged Corpus. The linguistic information identified by shallow syntactic parsing proves to be rich enough to support a number of large-scale NLP applications including information extraction (IE), text summarisation, machine translation, phrase identification in information retrieval, named entity identification, and a variety of text-mining operations. In order to achieve their goal, researchers of the University of Szeged, Department of Informatics, the MorphoLogic Ltd. Budapest, and the Research Institute for Linguistics at the Hungarian Academy of Sciences had to conduct some research concerning the syntax of Hungarian sentences, NP annotation schemes, and rules covering the recognition of phrases. Results showed that in Hungarian, nominal structures typically bear the most significant meaning (semantic content) within a sentence, therefore NP annotation seemed to be the most reasonable step forward.

Shallow parsing was carried out on the entire Szeged Corpus 2.0 (1.2 million words). Automated pre-parsing was completed with the help of the CLaRK[6] program, in which regular syntactic rules have been defined by linguistic experts for the recognition of NPs. Due to the fact that the CLaRK parser did not fully cover the occurring NP structures (its coverage was around 70%), manual validation and correction could not be avoided. In total, 250 thousand highest level NPs were found, and the deepest NP structure contained 9 NPs imbedded into each other. The majority of the hierarchic NP structures were between 1 to 3 NPs deep. Manual validation and correction lasted 60 person-months.

As a continuation of shallow parsing, the clause structure (CPs) of the corpus sentences was also marked. Labelling clauses followed the same approach as earlier phases of NLP: it comprised an automatic pre-annotation followed by manual correction and supplementation.

3 Use of the Szeged Corpus for training and testing machine learning algorithms

Due to the accurate and exhaustive manual annotation, the resulting corpus (both first and second versions) could serve as an adequate

[3] The Humor morpho-syntactic analyser is a product of the MorphoLogic Ltd. Budapest.

[4] The predicate $\texttt{after}_2$ e.g., denotes the second word to the right of the focus word.

[5] National Research and Development Programmes (NKFP) 2/017/2001 project funded by the Hungarian Ministry of Education, titled *Information Extraction from Short Business News*.

[6] The CLaRK system was developed by Kiril Simov at the Bulgarian Academy of Sciences in the framework of the BulTreeBank project.

database for the training and testing of machine learning algorithms. The applicability of these algorithms in Hungarian NLP was extensively studied in the past couple of years (Horváth et al., 1999), (Hócza et al., 2003). Researchers of the University of Szeged experimented with different kind of POS tagging methods and compared their results based on accuracy. Brill's transformation-based learning method (Brill, 1995) worked with 96.52% per word accuracy when trained and tested on the corpus. The HMM-based TnT tagger (Brants, 2000) performed 96.18%, while the RGLearn rule-based tagger (Hócza et al., 2003) produced 94.54% accuracy. Researchers also experimented with the combination of the different learning methods in order to increase accuracy. The best accuracy result, delivered by combining the above three methods, was 96.95%. Overall results showed that despite the agglutinating nature of Hungarian language and the structural differences between Hungarian and other Indo-European languages, all of the mentioned methods can be used effectively for learning POS tagging.

The applicability of machine learning methods for learning NP recognition rules was also investigated. The C 4.5 (Quinlan, 1993) and the RGLearn rule-based algorithms were selected for the learning process. NP recognition rules have been retrieved from the annotated corpus and were combined with manually defined expert rules. The main task of the NP recognition parser is to provide the best possible coverage of NP structures

Categories of recognition	**Precision**	**Recall**	**Accuracy**
Complete NPs	81.28%	87.43%	84.32%
Boundaries (first and last element) of NPs	88.31%	92.08%	90.54%
NPs (depth<=2)	86.02%	89.72%	88.37%
NPs (depth>2)	74.71%	78.19%	76.61%
Average	*82.58%*	*86.85%*	*84.96%*

Table 3. Test results of the RGLearn parser

The mentioned algorithms – although still under development – already perform between 80-90% accuracy (see Table 3.). Their performance strongly depends on the type of the processed text: phrase structures are recognised with better accuracy in news or technical type of texts than in student's compositions (where sentences are often grammatically inaccurate) or legal texts (where sentences are typically extremely long, and fragmented).

As a continuation of the work, an automated method was developed to perform IE from short business news. The 200 thousand word long, short business news section of the corpus was used as the training database for the IE tool. In the preparatory phase, the selected section of the corpus was enriched with semantic information. Possible semantic roles, such as SELLER, BUYER, PRODUCT, PRICE, DATE etc., were associated with each word, and were stored in a semantic dictionary. The most typical events of business life were represented by so-called semantic frames describing the relations of the different semantic roles. Possible frames were defined manually by linguists and allowed mapping between the lexical representation and the semantic role of a word. Semantic mapping rules were acquired by machine learning algorithms that used the manually annotated semantic roles as their learning source. The recognition of semantic frames was also supported by the series of NLP methods described earlier (i.e. POS tagging and shallow parsing).

During the developed information extraction process, the trained mapping tool takes a morpho-syntactically and syntactically annotated piece of text and performs two operations. First, it processes the morpho-syntactically disambiguated and shallow parsed text and assigns semantic roles to the words. The second operation determines relationships between the roles, i.e. maps semantic frames onto the existing structures. Semantic mapping is realised by simple pattern-matching methods using the frames previously defined by experts. Based on the results of the described operations, the mapping tool builds a semantic representation of the input text, already containing the required information. Results produced by this method were tested against the manually annotated corpus and showed that it identifies semantic roles with 94-99% accuracy and maps frames with up to 80% accuracy.

4 Current and future works

Current works aim at a more detailed syntactic analysis of the Szeged Corpus. With this, developers intend to lay the foundation of a Hungarian treebank, which is planned to be enriched with detailed semantic information as well in the future. The development of a suitable technique for the recognition and annotation of named entities (e.g., multi-word proper nouns) and special tokens (e.g., time expressions, dates, measures, bank account numbers, web- and e-mail addresses, etc.) is also planned in the near future. Further works aim at building firstly domain specific, later general ontologies and at developing automated methods that allow for extensive semantic analysis and processing of Hungarian sentences.

5 Related work

Corpus-based methods play an important role in empirical linguistics as well as in the application of machine learning algorithms. Annotated reference corpora, such as the Brown Corpus (Kucera, Francis, 1967), the Penn Treebank (Marcus et al., 1993), and the BNC (Leech et al., 2001.), have helped both the development of English computational linguistics tools and English corpus linguistics. Manual POS tagging and syntactic annotation are costly but allow one to build and improve sizable linguistic resources and also to train and evaluate automated analysers.

. The NEGRA (Skut at al., 1997) POS tagged and syntactically annotated corpus of 355 thousand tokens was the first initiative in corpus linguistics for German. The more recent TIGER Treebank project (Brants et al., 2002) aims at building the largest and most extensively annotated treebank for German. Currently, it comprises 700 thousand tokens of newspaper text that were automatically analysed and manually checked. Considerable results were achieved for Czech in the framework of the Prague Dependency Treebank project (Hajic, 1998), and for Bulgarian in the BulTreeBank project (Simov et al., 2003) as well.

The Szeged Corpus project is comparable both in size and in depth of analysis to the corpus and treebank initiatives mentioned above[7]. As the first such like initiative for Hungarian language, it is a valuable source for linguistic research and a suitable training and testing basis for machine applications and automated induction of linguistic knowledge.

References

Alexin Z., Csirik J., Gyimóthy T., Bibok K., Hatvani Cs., Prószéky G., Tihanyi L.: *Manually Annotated Hungarian Corpus* in Proc. of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), pp. 53-56, Budapest, Hungary (2003)

Brants, T.: *TnT - A Statistical Part-of-Speech Tagger,* in Proc. of the Sixth Conference on Applied Natural Language Processing (ANLP), Seattle, WA (2000)

Brants, S., Dipper, S., Hansen, S., Lezius, W. and Smith, G.: *The TIGER Treebank* in Proc. of the Workshop on Treebanks and Linguistic Theories, Sozopol, Bulgaria (2002)

Brill, E.: *Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging*, Computational Linguistics, vol. 21 (4), pp. 543-565, (1995)

Erjavec, T., Monachini, M.: *Specification and Notation for Lexicon Encoding*, Copernicus Project 106 „MULTEX-EAST", Work Package 1 – Task 1.1, Deliverable D1.1F (1997)

Hajic, J.: *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank* in Issues of Valency and Meaning, pp. 106-132, Charles University Press, Prague (1998)

Horváth T., Alexin Z., Gyimóthy T., Wrobel S.: *Application of Different Learning Methods to Hungarian Part-of-Speech Tagging* in Proc. of the 9th International Workshop on Inductive Logic Programming (ILP99), pp. 128-139, Bled, Slovenia and in the LNAI series vol 1634, Springer Verlag (1999)

Hócza A., Alexin Z., Csendes D., Csirik J., Gyimóthy T.: *Application of ILP methods in different natural language processing phases for information extraction from Hungarian texts* in Proc. of the Kalmár Workshop on Logic and Computer Science, pp. 107-116, Szeged, Hungary (2003)

Kucera H., and Francis, W. N.: *Brown Corpus Manual* Providence, Rhode Island, Brown University Press (1979)

Leech, G., P. Rayson and A. Wilson: *Word Frequencies in Written and Spoken English: based on the British National Corpus*, Longman, London (2001)

Marcus, M., Santorini, B., Marcinkiewicz, M.: *Building a large annotated corpus of English: the Penn Treebank* in Computational Linguistics, vol. 19 (1993)

Quinlan, J. R.: *C 4.5: Programs for Machine Learning,* Morgan Kaufmann Publisher (1993)

Simov, K., Simov, A., Kouylekov, M., Ivanova, K., Grigorov, I., Ganev, H.: *Development of Corpora within the CLaRK System: The BulTreeBank Project Experience* in Proc. of the Demo Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), pp. 243-246, Budapest, Hungary (2003)

Skut, W., Brants, T., Krenn, B., Uszkoreit, H.: *A linguistically interpreted corpus of German newspaper text* in Proc. of the Conference on Language Resources and Evaluation (LREC-98), pp. 705-711, Granade, Spain (1997)

[7] The different versions of the Szeged Corpus are available at http://www.inf.u-szeged.hu/hlt.