Evaluation of the NLP Components of an Information Extraction System for German

Thierry Declerck, Judith Klein, and Günter Neumann

German Research Center for Artificial Intelligence (DFKI GmbH)
Stuhlsatzenhausweg 3, 66123 Saarbrücken (Germany)
<Firstname.Lastname>@dfki.de

Abstract

This paper describes ongoing work on the evaluation of the NLP components of the core engine of smes (Saarbrücker Message Extraction System), which consists of a tokenizer, an efficient and robust German morphology, a part-of-speech (POS) tagger, a shallow parsing module, a linguistic knowledge base and an output construction component. Currently the morphology, the tagger and a parsing module (NP grammar) are under evaluation, at distinct degrees of progress. We present the methodology used and the results obtained so far.

1. Background

The integration of natural language processing (NLP) systems into a growing number of products, especially within the information technology, stimulates the consequent exploration of NLP functionalities for various purposes. The development of marked-oriented language technologies is greatly responsible for the increasing importance of the evaluation on NLP (sub-)systems. In the context of NLP based information extraction (IE) applications, for example, the reliable analysis of natural language input is a necessary prerequisite to ensure high quality results of the IE task. And only profound diagnostic evaluation cycles during system development and adequacy evaluations on prototypes can provide information on the performance of the NLP functionalities and direct application-oriented improvements.

But several aspects make evaluation on NLP systems a difficult task. One core issue frequently mentioned as an important precondition for profound evaluation is the availability of suitable language data as test and reference material. While linguistic competence data, e.g. systematically constructed test suites, typically serve for diagnostic evaluations, annotated corpora mainly aim to test the performance of NLP systems with respect to a particular application. Since the building of annotated language resources is a laborious and time-consuming task, re-usability as well as automatic annotation are key issues in the area of NLP evaluation. This background influenced the evaluation work described below.

2. Introduction

In this paper we present the ongoing evaluation work of the NLP components of smes (Saarbrücker Message Extraction System), developed at the DFKI (see Neumann *et al.*, 1997). One of the goal followed with the development of smes is to provide a parameterizable core machinery for a variety of commercially viable applications in the area of information extraction (IE). This goal is currently being pursued within the PARADIME (Parameterizable Domain-adaptive Information and Message Extraction) project[1].

The smes system is modularly designed and consists of a tokenizer, an efficient and robust German morphology, a part-of-speech (POS) tagger, a shallow parsing module, a linguistic knowledge base and an output construction component. The current evaluation study (whose design - taking into account former evaluation methodology findings - is presented in section 3) is concerned with the progress evaluation, under glass-box condition, of the NLP components on the basis of a annotated reference corpus taken from a German business magazine. We report on the first results on the evaluation of the morphological analyzer and the part-of-speech tagger as well as on the evaluation setting for the parsing module, which is both test-suite-based and corpus-based. It is important to note, that for the time being the study is concerned only with the NLP components of smes and not with its IE functionality.

Since PARADIME is concerned with the development of a core machinery for a variety of IE applications, the project will also investigate to what extent test data can easily be provided for each application domain, allowing thus the evaluation of the derived IE system for the specific applications. For this purpose we examine to what extent some NLP components of smes can be used to support a semi-automatic annotation of test corpora for specific domains. This fact provides a second good reason for a profound evaluation of the NLP modules envisaged for this annotation task (i.e. morphological and POS tagging), since those components must guarantee that the analysis results subsequently used for annotation are reliable and need only minor manual post-editing. In this context we are also evaluating the use of an adapted version of the unsupervised Brill tagger (Brill, 1995) for German to support the semi-automatic POS tagging of reference corpora.

3. Evaluation design

In a former project - the COSMA (Cooperative Schedule

[1] See http://www.dfki.de/lt/projects/paradime-e.html.

Management Agent) project[2] - a first prototype of smes was integrated as the (shallow) syntactic component of a German dialogue system for appointment scheduling via e-mail. The evaluation study of smes as the analysis component of this application (see Klein *et al.*, 1997) revealed important findings which influenced the evaluation design of the components of smes for the PARADIME project:

Setting The evaluation of the final linguistic output of smes made it difficult to identify and localize the system's internal deficiencies. Within the current progress evaluation, all processing parts are being tested in isolation under glass-box conditions.

Reference Material The lack of suitably annotated language resources hampered the examination and interpretation of the system results. Within PARADIME a reference corpus is being built which is augmented with exactly that annotation suitable for the component under evaluation. Additionally to the corpus-based approach, the parsing modules will be checked on suitable syntactic test suite data.

Testing The checking of the smes results was done manually and thus a laborious and time-consuming task. On the basis of annotation schema developed for PARADIME the (semi-) automatic comparison between the reference corpus and the output of the smes components will be supported.

Measurement The actual smes output was immediately assigned a quality predicate which lead to quite subjective judgments. For the new evaluation scenario, there is an intermediate step calculating the row output of the smes components in numeric terms, interpreted as *recall* and *precision*.

4. Language Resources

The provision of suitable language resources is an important prerequisite for profound and reliable evaluation studies. But even though the importance of test and reference material is widely acknowledged the time and costs which must be invested, mostly prohibit the construction of large and richly annotated data pool. Only if the annotated resources can be of use for other than a specific project or system, is the investment justified. Another way of coping with this problem is to investigate the development of techniques for the efficient (semi-)automatic annotation of language data. This point is at some places addressed in this paper, but reliable results can be provided first after a long-term study within a project concerned directly with annotation strategy and techniques, which is not the case for PARADIME, where the annotation issue is discussed in the context of the accompanying evaluation study of the advanced development of an NLP system.

4.1. A Reference Corpus for PARADIME

An important aspect of the improvement of the IE-core system is the evaluation of the performance of its components applied to large texts. For this purpose, a corpus from the German business magazine "Wirtschaftswoche" from 1992 (1.4 Mega-Byte), consisting of 366 texts on different topics written in distinct styles, was selected. Augmented with the necessary annotation, these texts will serve as a reference basis to support the evaluation of the NLP components of smes.

Since the manual building of a large, richly annotated reference corpus is cost and time intensive, the components of smes will be employed to support the annotation work. So the whole corpus has been automatically marked up with lexical information provided by the morphology component of smes, where unknown words are assigned the tag N_V_A (noun, verb or adjective).

A large part of the ambiguous words has been manually disambiguated and now more than 80% of the morpho-syntactically annotated corpus has been validated. The use of an unsupervised tagging strategy for supporting the semi-automatic POS tagging of the rest of the reference corpus is under investigation.

We are trying to make the morpho-syntactically annotated reference corpus reusable for other projects and we defined amongst others a mapping to the Stuttgart-Tübingen Tag-Set (STTS), which is widely acknowledged in the German speaking Computational Linguistics community (see Thielen and Schiller, 1996).

The validated part of the corpus is currently being tagged with structural information, including agreement. We started with the nominal phrases, considering both the phrasal level and the internal structure in terms of head-modifier relations. Up to 20% of the corpus has been tagged with respect to the constituent structure and agreement information has been added to the mother node of the NPs. Furthermore, we will investigate if and how the corresponding parsing module of smes can help to semi-automatically annotate the rest of the corpus with respect to NPs, or at least to certain sub-classes of nominal constructions.

The corpus will be further annotated wrt. clause level valency information in order to allow an evaluation of the first prototype of a recently implemented algorithm for the extraction of grammatical functions.

Thus, the same resource will serve as a reference corpus for the distinct evaluations of the NLP components of smes.

4.2. A Test Suite for PARADIME

The evaluation of the parsing module will not only be corpus-based but will additionally exploit the TSNLP (c.f. Lehmann

[2] See Busemann *et al.*, 1997. For more information on this project, see http://www.dfki.de/lt/projects/cosma.html.

et al., 1996)[3] and DiET (Netter *et al.*, 1998)[4] resources for the building of suitable test suites for the syntactic component of smes. The German TSNLP test suite provides not only richly annotated test sentences but also a reasonable amount of complex nominal phrases which will be relevant for the testing of the NP grammar. Based on the comprehensive TSNLP guidelines for the construction and annotation of test suites additional data can easily be supplied for all parsing modules. It is envisaged to employ the sophisticated annotation tool which is currently being developed within DiET to support the syntactic annotation of the test suite data.

5. SMES – Saarbrücker Message Extraction System

The core engine of smes consists of a tokenizer, an efficient and robust German morphology, a POS tagger, a shallow parsing module, a linguistic knowledge base and an output construction component. These components are briefly presented below.

The tokenizer The tokenizer, implemented in lex, is based on regular expressions, against which the textual input is matched for identifying some text structures (e.g. words, date expressions, etc.). Number, date and time expressions are normalized and represented as attribute value structures.

MORPHIX++ The morphological analysis is performed by MORPHIX++, an efficient and robust German morphological analyser which performs morphological inflection and compound processing[5]. MORPHIX++ has a very broad coverage (a lexicon of more than 120,000 stem entries). The output of MORPHIX++ consists of a list of all possible readings, where each reading is uniformely represented as a triple of the form <STEM,INFLECTION,POS>.

The Brill Tagger Ambiguous readings delivered by MORPHIX++ are disambiguated wrt. POS using an adapted version of Brill's unsupervised tagger (Brill 1995) for German, which is fully integrated in the system architecture of smes[6]. The integrated BRILL tagger consists of a learning component and an application component. Within smes the learning material for the tagger consists of the lexical mark-up of texts provided by MORPHIX++. In the learning phase the tagger induces disambiguation rules on the basis of the morpho-syntactic annotated texts. In a second phase those rules are applied to the morphologically analysed input texts. In PARADIME the use of the Brill's unsupervised tagging strategy is also investigated in order to see to what extent it can be dispensed with the time consuming building of training corpora for every specific application[7].

The parsing module A shallow parsing component consisting of a declarative specification tool for expressing finite state grammars. Shallow parsing is done in two steps: Firstly, specified finite state transducers (FST) perform fragment recognition and extraction on the basis of the output of the scanner and of MORPHIX++. Fragments to be identified are user-defined and typically consist of phrasal entities (e.g. NPs, PPs) and application-specific units (e.g. complex time and date expressions). In the second phase, user-defined automata representing verb frames operate on the extracted fragments to combine complements and adjuncts into predicate-argument structures.

The linguistic knowledge base The knowledge base includes a lexicon of about 120,000 root entries and various sub-grammars (for complex time and date expressions, general NPs etc.). The design of the lexicon allows for structured extensions. So for example, information about sub-categorization frames - extracted from the Sardic lexical data base (see Buchholz, 1996) developed at the DFKI - has been recently integrated into the verb lexicon. The system now has 11998 verbs with a total of 30042 subcategorization frames at its disposal.

The output construction component An output construction component is using fragment combination patterns to define linguistically head-modifier constructions. The distinction between this component and the automata allows a modular I/O of the grammars. The fragment combiner is also used for the instantiation of templates, providing one of the possible visualization of the results of the IE task.

6. Evaluation of NLP components of smes

The NLP components of smes are under progress evaluation, following their ordering in the processing chain of smes (but the tokenizer was not taken into consideration).

6.1. Evaluation of MORPHIX++

Since the output of MORPHIX++ is providing the annotation basis for the reference corpus, the evaluation started with several cycles of diagnostic evaluation in order to determine the linguistic coverage and detect shortcomings of the morphological component. The first test run on the whole corpus showed that MORPHIX++ had a lexical coverage of 91,12%. The systematic inspection of the morphological mark-up revealed some necessary extensions and modifications concerning missing lexical entries, incomplete or

[3]The EU project TSNLP (test suites for natural language processing) ended in 1996 and information on the project and its results can be obtain via http://tsnlp.dfki.uni-sb.de/tsnlp/

[4]The EU project DiET (diagnostic and evaluation tools for natural language applications) started in March 1997. Information is available under http://dylan.ucd.ie/DiET/.

[5]MORPHIX++ is based on MORPHIX (Finkler and Neumann, 1988).

[6]First experiments done with the Brill's unsupervised tagger within smes were very promising and Neumann *et al.* (1997) reported already a tagging accuracy as high as 91.4%.

[7]In the future we might go for combining the supervised and unsupervised strategies. The supervised Brill tagger is currently being trained for German at the University of Zürich; see http://www.ifi.unizh.ch/CL/tagger

erroneous single word analyses and some general aspects of the assignments of morpho-syntactic information. This examination task was supported by some tools. So for example, a tool was delivering an ordered listing of all the words, classifying them into categories or classes of ambiguities. This allowed words having a very high frequency in the corpus to be checked just once. The current coverage of MORPHIX++ has reached about 94%, where most of the words not analyzed are, in fact, proper nouns or misspelled words.

6.2. Evaluation of the POS Tagger

After a first correction phase on MORPHIX++, the BRILL tagger was run over the annotated corpus using a first version of a simple rule application strategy. The context taken into account for disambiguation consisted just of one word or one category on each side of the processed item.

While in this first evaluation round the input texts are annotated with POS information only, in the next steps additional features (e.g., case) will be considered until the best feature set for optimal disambiguation results can been determined. For the realization of this experiment, we will take advantage of the recently implemented parameterizable interface of MORPHIX++, defining a cardinality of possible tag sets which ranges from 23 (considering only POS) to 780 (considering all possible morpho-syntactic information).

The evaluation of the BRILL tagger is based on quantitative and qualitative measures. The *recall* is calculated as the ratio between the number of disambiguations performed and the number of all word ambiguities provided by MORPHIX++. The *precision* of the disambiguation step is measured as the ratio between the number of correct disambiguations performed and the number of all words of the manually validated reference corpus (since they are supposed to be annotated with the correct reading). A set of tools was implemented to support the automatic comparison and calculation of the output of the BRILL tagger and the MORPHIX++ output and the reference material respectively. The following results were reported: 62% of the ambiguities in the input text were disambiguated showing an accuracy of 95%. These results are promising since they have been obtained on the basis of a basic implementation of the BRILL tagger.

Even though it is not yet clear which recall and precision values are necessary to express reliable results, the tagger's current performance is certainly not sufficient for the semi-automatic annotation of the reference corpus. But we investigate the manual integration of additional linguistically motivated disambiguation rules as well as the addition of a small amount of morpho-syntactically language data to the input. We expect considerable improvements of the recall values of the tagger.

We also expect further improvements of the performance of the unsupervised BRILL tagger through a generalisation of the training phase. This generalisation enables the learning of complex rules (like "change X to Y when one of the three preceding words has category Z" or "Change X to Y when the preceding word has category Z AND the following word is W"). The testing of distinct parameterization of the tagger is on the way (Becker, 1998). First results on the basis of small-scale experiments are very promising and a considerable improvement of the performance can be expected. But reliable results can be presented only after a large-scale experiment covering the whole reference corpus has been carried out.

From the actual evaluation study we also expect to gain valuable information about the necessary amount of annotated text to support and control the performance of lexical tagging.

6.3. Evaluation of Shallow Parsing

Since the performance of the parser partly depends on the output of MORPHIX++ and the BRILL tagger, high quality output of the preprocessing steps will positively influence the parsing results. In order to check and improve the functionality of the isolated syntactic analysis, diagnostic evaluations under glass-box conditions are necessary.

The design of the evaluation of the NP grammar (and later other grammar modules) follows two strategies, the first one being corpus-based and the second one being test-suite-based. The test-suite-based evaluation is necessary, since one can not be sure that the corpus will contain all the relevant phaenomena. And also the competence of the grammar can be more systematically evaluated. On the other side, the corpus-based approach can reveal phaenomena which have not been included so far in the test suite.

In the context of the corpus-based approach, up to 20% corpus has been annotated with structural information, i.e. all simple and complex NPs are marked and augmented with agreement information (attached at the mother node). The test data will also contain annotations describing the internal head-modifier dependencies within the phrases. First test runs to check the recognition functionality of the NP grammar will be carried out. At the same time, the corpus NPs will be annotated with head-modifier information – possibly in the same format as delivered by the parsing module – to represent the internal structure of the nominal constructions. On the basis of the syntactically interpreted reference corpus, further evaluations will be carried out in order to check the accuracy of the sub-grammar. Provided the annotation format and the parsing output are identical, automatic comparison routines will be employed to support the evaluation procedure.

For the purpose of the test-suite-based evaluation, the TSNLP test suite for German, which comprises a substantial amount of annotated test items exemplifying various syntactic phenomena, is exploited.[8]. Test data on complex NP modification and word order phenomena within NPs can be supplied by the TSNLP data base. Additional prototypical items are being constructed in order to account for a wide range of NP constructions. In this evaluation phase the test items will be fed one-by-one to the NP automata and the output will be checked manually. Since the NP test suite will probably not contain more than a hundred test items and the results can easily be examined and interpreted by the system developers, this manual procedure can

[8]Test data can be retrieved from the database under http://tsnlp.dfki.uni-sb.de/tsnlp/tsdb/tsdb.cgi.

be justified.

The same methodolgy will be applied for the other parsing modules of smes. When these evaluation parts have been completed an overall study considering just the final parser output will be carried out.

7. Conclusions and Future Work

We have presented the progress evaluation of NLP components accompanying the development of smes as an adaptive core engine for a variety of IE applications.

The applied evaluation follows two strategies; (i) a corpus-based evaluation, for which a reference corpus was established (investigating as well efficient annotation strategies) and (ii) a test-suite-based evaluation, based partly on the work done within the TSNLP project (and investigating a closer integration of results of the ongoing DiET project).

We also investigate if for the evaluation of the adapatation of smes for specific applications a fast way of providing relevant reference corpus can be realised. The first result of the evaluation of the morpho-syntactic and POS tagging components of smes for their use as supporting tools for annotation are very promising wrt. this issue.

In the future we will also start the evaluation of smes as a core engine for IE tasks. For this we will adapt another evaluation strategy and go there for a blind scientific assesment of system output, under black box conditions, looking for some synergy with the research work done on evaluation of IE systems within the MUC initiative (sse Grishman and Sundheim 1996).

8. Acknowledgements

The research underlying this paper was supported by grants from the German Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMB+F) for the DFKI project PARADIME (FKZ ITW 9704). We would like to thank Milena Valkova and Birgit Will for their patient and precise annotation work, and Markus Becker for his work on the integration of the BRILL tagger.

9. References

Becker, M. (1998). Unsupervised Part of Speech Tagging with Extended Templates. To appear in Proceedings of the Student Seesions of the 10th ESSLLI, 1998.

Brill, E. (1995). Unsupervised learning of dismabiguation rules for part of speech tagging. Proceedings of the Second Workshop on Very Large Corpora, WVLC-95 Boston, 1995.

Buchholz, S. (1996). Entwicklung einer lexikographischen Datenbank für die Verben des Deutschen. Master Thesis, Universität des Saarlandes.

Busemann, S.; Declerck, T.; Diagne A.K.; Dini, L; Klein, J. Schmeier, S. (1997). Natural Language Dialogue Service for Appointment Scheduling Agents. Proceedings of the 5th Conference on Applied Natural Language Processing, ANLP-97, pp.25-32, Washington DC, 1997.

Finkler, W.; Neumann, G. (1988). Morphix: A fast Realization of a Classification–based Approach to Morphology. Proceedings der 4. Österreichischen Artificial Intelligence Tagung, Wiener Workshop Wissensbasierte Sprachverarbeitung, pp.11-19, Berlin, 1988.

Grishman, R.; Sundheim, B. (1996). Message Understanding Conference – 6: A Brief History. Proceedings of the 16th International Conference on Computational Linguistics, COLING, 1996.

Klein, J.; Busemann, S.; Declerck,T. (1997). Diagnostic Evaluation of Shallow Parsing Through an Annotated Reference Corpus. Proceedings of the Speech and Language Technology Club Workshop, SALT-97, pp.121-128, Sheffield, 1997.

Klein, J.; Declerck, T.; Neumann, G. (1998). Evaluation of the Syntactic Analysis Component of an Information Extraction System for German. Proceedings of the Workshop on The Evaluation of Parsing Systems, Granada, 1998.

Lehmann, S.; Oepen, S.; Estival, D.; Falkedal, K.; Compagnion,H.; Balkan, L.; Fouvry, F.; Arnold, D.; Dauphin, E.; Régnier-Prost, S.; Lux, V.; Klein, J.; Baur, J.; Netter, K. (1996). TSLNP – Test Suites for Natural Language Processing. Proceedings of the 16th International Conference on Computational Linguistics, COLING-96, pp. 711-126, Copenhagen, 1996.

Neumann, G.; Backofen, R.; Baur, J.; Becker, M.; Braun, C. (1997). An Information Extraction Core System for Real World German Text Processing. Proceedings of the 5th Conference on Applied Natural Language Processing, ANLP-97, pp.209-216, Washington DC, 1997.

Netter, K.; Armstrong, S.; Kiss, T.; Klein, J.; Lehmann, S.; Milward, D.; Petitpierre, D.; Pulman, S.; Regnier-Prost, S.; Schäler, R.; Uszkoreit, H.; Wegst, T. (to appear). DiET - Diagnostic and Evaluation Tools for Natural Language Applications. Proceedings of the first international Conference on Language Resources and Evaluation. Granada, Spain, 1998.

Oepen, S.; Klein, J.; Netter, K. (1997). TSLNP – Test Suites for Natural Language Processing. In: Nerbonne, J. (Ed.). Linguistic Databases. University of Stanford, CSLI Lecture Notes, Stanford, pp.13-37, 1997.

Thielen, C.; Schiller, A. (1996). Ein kleines und erweitertes Tagset fürs Deutsche. In Feldweg, H.; Hinrichs, E. (Eds)., Lexikon & Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen., pp. 193-203. Tübingen: Niemeyer, 1996.