DGfS-CL

Maria-Hendrike Peetz
Universiteit van Amsterdam

Using Suffix Trees for Text Categorization in Computational Linguistics

General similarity measures, within the domain of computational linguistics, are an important tool to distinguish words, sentences, texts, and documents. Probabilistic generalized suffix trees (PGST), as commonly used in the field of bioinformatics for DNA sequences, are a natural candidate for computational linguistics. I have developed a successful similarity measure for text classification. The underlying mechanism uses a count of the shared nodes in a common PGST, applying different weighting factors. To further account for text length, normalization is performed with different tree properties, for example density, vocabulary, and tree size.

From a language modeling point of view this method successfully evaluates inter-document similarity. It analyses interpolated n-grams similar to [1]. Additionally, deep trees contain information about long-distance relationships of words or part-of-speech tags used in the texts.

The current version of the algorithm incorporates the similarity measure into a k-nearest neighbor classifier, to analyze the impact of different suffix tree based similarity calculations on the results. The experimental corpus included "raw" texts as well as part-of-speech tags featuring a mixed representation of both.

The results were astonishing, a 92% accuracy for a gender classification task, using the British National Corpus. I can further report stability regardless of tree depth, which is required for a fully automated algorithm. This is to say, no manual feature selection is required.

[1] R. Kneser. Statistical Language Modeling using a Variable Context length. In *Proc. ICSLP '96*, volume 1, pages 494–497, Philadelphia, PA, 1996.

DGfS-CL

Fabienne Fritzinger
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Azenbergstr. 12
70174 Stuttgart

Extraction of German Multiword Expressions based on Parallel Text

The poster deals with the automatic extraction of German Multiword Expressions (MWE) by applying statistical methods on a multilingual, parallel corpus. It is an extension of previous work by [1].
The focus is on the extraction of German verb + PP combinations of the kind *ins Leben rufen* ("to initiate", lit.: "to call into life"). To extract such MWEs, their idiosyncratic (i.e. non-compositional) semantics is exploited. It is assumed that opaque combinations are translated as a whole, whereas compositional uses would show regular, individual translations of the words involved. The required translations are obtained via word alignment (GIZA++ [2]) on the EUROPARL corpus (~ 1.5 milion sentences per language [3]).
The extraction procedure consists of three stages: 1) creating a candidate list, 2) acquisition of the candidates' translations, and 3) ranking the candidate list. To get a candidate list of verb + PP combinations, the German section of EUROPARL is dependency-parsed [4]. Then, two *link lexicons* are established: the global link lexicon contains single word translations across the whole corpus, whereas the local link lexicon contains only translations of candidate MWEs. Based on these link lexicons, two statistical measures, namely the *translational entropy* and the *proportion of default alignments,* are calculated to rank the candidates in decreasing probability of being a valid MWE.
Numerous experiments are performed to further optimise the original method. The impact of different individual parameters (e.g. a restricted verb list, data from more than one language pair, a preference for adjacent MWE candidates) as well as combinations of these parameters on result quality has been investigated. This leads to the following results: depending on the actual parameter settings, *uninterpolated average precisions* between 0.908 and 0.988 are reached.
Further investigations deal with the effects of considerably reducing the amount of linguistic knowledge in the extraction procedure.

References

[1] Villada Moirón, B., Tiedemann, Jörg; Identifying idiomatic expressions using automatic word alignment. In: Proceedings of the EACL 2006 workshop on multiword expressions in a multilingual context; 2006.
[2] Och, F.J.; Ney, H.; A systematic comparison of various statistical alignment models. Computational Linguistics 29 (1); 2003.
[3] Koehn, P.; Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit 2005.
[4] Schiehlen, M.; A cascaded finite-state parser for German. In: Proceedings of the 10th EACL; 2003.

DGfS-CL

Jaouad Mousser
Universität Konstanz

A VerbNet for Arabic

Abstract

Verb Lexica like VerbNet and FrameNet have become an essential part of many computer linguistic applications like text summarization, question-answering etc. For example: a question-answering system like Parc-Bridge 1 from PARC XEROX is based (among other things) on the precise representation of semantic informations from verbs and the mapping of argument structures of the parsed sentences in the appropriate thematic structures. In the center of the system stands a version of VerbNet for English converted in a PalmDataBase (pdb).
To develop similar systems for few focused languages, computer linguists may encounter the problem of lack of resources. This applies particularly to not-European languages like Arabic.
For my question-answering system for arabic, which is essentially inspired by Parc-Bridge, a handcrafted annotation of verbs with their corresponding argument and thematic structures was crucial. The obtained annotated collection of 2500 verbs from deferent new papers like Al jazeera, Alquds al arabi , Daralhayat and Almassae provoks in addition the idea of classifying them into Palmer verb classes. After some experiences, it turns out that the annotated collection can be converted with few efforts to a VerbNet similar lexicon.
In my Poster I present a VerbNet similar lexicon for the modern standard-Arabic (MSD). The verbs are annotated with mmax2 and stored in XML files according to the above mentioned functions. With a Java applet with a XML query function, the verb database is queried for verb argument - and thematic structures and the mapping between them. In adition I will compare the classes of some arabic verbs with their correponding english verbs concerning phenomena like causativity, reflexivity, multi-thematic structures etc..

References

---
http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
http://www.eml-research.de/english/research/nlp/download/mmax.php

DGfS-CL

Christian Chiarcos*, Thomas Krause+, Anke Lüdeling+, Julia Ritz*, Viktor Rosenfeld+, Manfred Stede*, Amir Zeldes+ and Florian Zipser+

* Universität Potsdam
+ Humboldt-Universität zu Berlin

Search and Visualization of Richly Annotated Corpora with ANNIS2

This poster presents the latest version of ANNIS2, a web browser-based search and visualization environment designed to access richly annotated corpora with heterogeneous annotation schemes. Developed within Collaborative Research Centre 632 (SFB 632: "Information Structure: The Linguistic Means for Structuring Utterances, Sentences and Texts"), ANNIS (ANNotation of Information Structure) must meet the requirements imposed by diverse data from partner projects within the Research Centre and beyond.

Since information structure interacts with linguistic phenomena on many levels, the need to concurrently query and visualize data annotated for syntax, semantics, morphology, prosody, phonetics, referentiality and lexis, must be addressed, including where the data is multimodal. For this reason, ANNIS2 supports annotations of tokens, token spans and trees or other DAGs (directed acyclic graphs), and uses an appropriate query language capable of searching these structures. Both query language and visualizations are fully Unicode compatible to ensure support for a wide variety of non-European languages.

The underlying data for the system is annotated using both automatic taggers/parsers and a small set of manual annotation tools: EXMARaLDA (Schmidt 2004), annotate (Brants & Plaehn 2000) / Synpathy (www.lat-mpi.eu/tools/synpathy/), MMAX2 (Müller & Strube 2006), RSTTool (O'Donnell 2000) and PALinkA (Orasan 2003). These are then mapped onto the encoding standard of the SFB, PAULA (Potsdamer AUstauschformat für Linguistische Annotation / Potsdam Interchange Format for Linguistic Annotation), a stand-off multilevel XML format, which serves as the basis for further processing. The XML data is compiled into a relational database scheme, making the system's backend particularly scalable.

References

Brants T. & Plaehn, O. (2000) Interactive Corpus Annotation. In: *Proc. LREC 2000*, Athens.

Müller, C. & Strube, M. (2006), Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun, S., Kohn, K. & Mukherjee, J. (eds.), *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, 197–214.

O'Donnell, M. (2000) RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory. In: *Proc. of the International Natural Language Generation Conference (INLG'2000), 13-16 June 2000*, Mitzpe Ramon, Israel, 253–256.

Orasan, C. (2003), Palinka: A Highly Customisable Tool for Discourse Annotation. In: *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo.

Schmidt, T. (2004) Transcribing and Annotating Spoken Language with Exmaralda. In: Proceedings of the LREC-workshop on XML Based Richly Annotated Corpora, Lisbon 2004. Paris: ELRA.

UNIARAB: AN UNIVERSAL MACHINE TRANSLATOR SYSTEM FOR ARABIC BASED ON ROLE AND REFERENCE GRAMMAR

Yasser Salem and Brian Nolan
Department of Informatics
Institute of Technology Blanchardstown, Dublin, Ireland
E-mails: {firstname.surname}@itb.ie

ABSTRACT

This poster reports on work-in-progress, with details on significant research results, on our investigations into, and the development of, a rule-based lexical framework for Arabic natural language processing using the Role and Reference Grammar (RRG) linguistic model (Van Valin and LaPolla 1997, Van Valin 2007)). A system called UniArab is introduced in this research to support the framework.

We outline the conceptual structure of UniArab System, which utilizes the framework and machine translates the Arabic language into another natural language, in this instance English. Also, we explore how the particular linguistic characteristics of the Arabic language (including syntax and word morphology) will effect the development of a Machine Translation (MT) tool (Hutchins 2003, Izwaini 2006) from Arabic to English. Several distinguishing features of Arabic (Ryding 2005) pertinent to MT will be explored in detail with reference to some potential difficulties that they might present.

We use the RRG theory to motivate the architecture of the lexicon and the RRG bidirectional linking system to design and implement the parse and generate functions between the syntax-semantic interfaces. Our research has yielded significant results and in comparison with translations of (source) modern standard Arabic sentences in the native orthography from other software systems, we deliver more accurate and grammatical output in (target) English. This is, we believe, to the use of the RRG linguistic model in software.

Hutchins, W. J. 2003. *Machine translation: General overview*. Oxford: Oxford University Press.

Izwaini, S. 2006. *Problems of Arabic machine translation: evaluation of three systems*. *The British Computer Society (BSC), London, pages 118–148.*

Ryding, K. C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Salem, Y., Hensman, A., and Nolan, B. 2008. *Towards Arabic to English machine translation.* ITB Journal May 2008 Issue 17: accessible at: http://www.itb.ie/site/researchinnovation/itbjournal.htm

Van Valin, R. 2007. *The Role and Reference Grammar analysis of three place predicates*. Accessible at: http://linguistics.buffalo.edu/people/faculty/vanvalin/rrg.html

VanValin, R. and LaPolla, R. 1997. *Syntax: Structure, Meaning, and Function*. *Cambridge: Cambridge University Press.*

DGfS-CL

Victoria Oketch
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin

Multilingualität und Lernerkorpora

In meinem Poster geht es um die Frage, inwieweit Lernerkorpora zur Erforschung der unterschiedlichen Einflüsse verschiedener bereits gelernter Sprachen auf die zu erlernende Fremdsprache Deutsch dienen können.

Lernerkorpora haben sich in den letzten Jahren als eine gute Datengrundlage für viele Fragen der Erforschung von Spracherwerbsverläufen etabliert. Dabei gibt es viele Studien zum Einfluss der jeweiligen Muttersprache auf den Erwerbsverlauf in der Fremdsprache. Viele Lerner sind allerdings nicht bilingual, sondern multilingual (wobei es unterschiedliche chronologische Verläufe und Gebrauchskontexte der bisher gelernten Sprachen gibt), und die L_n haben jeweils Einflüsse auf die neu zu lernende Sprache L_{n+1}.

Die Einflüsse der jeweiligen verschiedenen Sprachen auf die neue und zu erlernende Sprache können unter Berücksichtigung bestimmter Bedingungen untersucht werden. Jedoch müssten diese Bedingungen auch im afrikanischen Sprachenkontext Geltung tragen.

Ich habe im Rahmen des Falko-Projekts

(http://www2.hu-berlin.de/korpling/projekte/falko/index.php) in Kenia, einem Land mit einer hohen Mehrsprachigkeitsquote, ein Essaykorpus erhoben und diskutiere, wie die Einflüsse der jeweiligen L_n ermittelt werden können.

Marc Emmerich
FSU Jena
Lehrstuhl für Indogermanistik
Projekt Historische Syntax des Jiddischen

Tacheles für die Annotation Jiddischer Texte

Das Programm Tacheles ist ein Annotationswerkzeug zur manuellen und automatischen Annotation sprachlicher Merkmale auf mehreren Ebenen. Es wird im Rahmen des DFG-Projektes zur Historischen Syntax des Jiddischen (HSJ) entwickelt. Es unterstützt den Benutzer einerseits mit einer intuitiven und effizienten Steuerung, andererseits können Sprachdaten automatisch aufbereitet werden.
Die Steuerung erlaubt den gleichzeitigen Einsatz von Maus und Tastatur. Sie passt sich dynamisch an das jeweils ausgewählte Annotationsziel an: Der Benutzer kann nur Tags aus dem aktuell verwendeten Tagset auswählen. Zudem werden logische Widersprüche im Vorhinein ausgeschlossen. Beispielsweise kann für das Wort „gehen“ kein Genus annotiert werden, da es sich um Verb handelt. Für das Wort „Garten“ kann Genus annotiert werden, Person jedoch nicht. Tacheles dient jedoch nicht nur der Annotation. Es unterstützt den Benutzer auch beim Vergleich und bei der Zusammenführung von Corpora. Insbesondere lässt sich im Programm selbst die Leistungsfähigkeit des automatischen Parsers ermitteln.
Zur automatischen Annotation greift Tacheles auf eine Wörterbuchdatenbank zurück, die im Laufe des HSJ-Projektes gefüllt wird. Sie enthält neben den Wörtern und deren zugeordneten Merkmalen auch statistische Informationen, die der Auflösung mehrere Varianten dienen. Zusätzlich werden die statistischen Daten von dem integrierten POS-Tagger verwendet. Er ermittelt auf der Basis eines Hidden-Markov-Modells auch bisher unbekannte, aber statistisch wahrscheinliche Zuordnungen.
Tacheles ist vollständig in Java programmiert, alle Daten – von den Corpora über die Annotationen zu den verwendeten Tagsets - sind in einem TEI-konformen XML-Format gespeichert. Die damit erreichte Plattformunabhängigkeit und die vergleichsweise leichte Erweiterbarkeit sollen das Programm über den Bereich des HSJ-Projektes hinaus interessant machen.
Die Speicherung im XML-Format bringt den weiteren Vorteil, dass die Daten mittels XSL-Translatoren verarbeitet werden können. Dies ermöglicht die bequeme Absuche des Corpus mit Hilfe gängiger Webbrowser.

DGfS-CL

Thomas Mayer
Universität Konstanz

Unsupervised Syllabification

The present work describes the automatic syllabification of words on the basis of the distribution of possible word-initial onsets and word-final codas in a reasonably-sized corpus of the language under consideration. The algorithm is totally unsupervised, i.e., it does not require any language-specific knowledge but is designed to work for all natural languages provided their copora are in reasonably good phonological transcription (as is the case with most languages that only recently adopted the Latin alphabet).

The program works as follows. First, it makes use of Sukhotin's algorithm (Sukhotin 1962) to distinguish vowels and consonants in the input text. Then, the initial and final consonant clusters of each word (delimited by the left word edge and the first vowel as well as the last vowel and the right word edge, respectively) are used as an approximation of possible word-internal consonant clusters. The assumptions are that vowels mark syllable peaks and that internal clusters can be separated on the basis of the frequency of occurrence of the constituting parts at word edges. There have been several suggestions in the literature on how the syllable boundary for those word-internal consonant clusters can be determined (see Vogel 1977 for an overview). Some of these have been implemented for the online version of the program (`http://typo.uni-konstanz.de/syll/`) and can be selected from a menu. Sequences of two vowels are treated differently depending on whether both vowels occur in the corpus more often next to each other (in which case they are assumed to be tautosyllabic) or separated by a consonant (making them heterosyllabic).

The algorithm has been tested on a varierty of geographically and genetically diverse languages. The results as well as the program itself will be presented.

References:

Sukhotin, Boris V. 1962. Eksperimental'noe vydelenie klassov bukv s pomoščju EVM, *Problemy strukturnoj lingvistiki*, volume 234, 189-206.

Vogel, Irene. 1977. *The Syllable in Phonological Theory with Special Reference to Italian*, Stanford University.

DGfS-CL

Stefanie Dipper / Heike Zinsmeister
Ruhr-Universität Bochum / Universität Konstanz

Feature Selection for Local Coherence

Local coherence, the way adjacent sentences are glued together, arises out of a number of different factors. Local coherence results, e.g., from *syntactic* similarity (like parallelism) between adjacent sentences. Similarity or relatedness at the *semantic* level is even more important for local coherence, e.g. in the form of coreferent expressions or lexically-related words. *Information-structural* features and *discourse* relations also play a role for local coherence. Centering Theory (CT, Grosz et al. 1995) and the entity-based approach by Barzilay&Lapata (2008) model local coherence by reference to the syntactic functions of coreferent expressions, or, in a CT version adapted to German, by reference to information-structural features (givenness/familiarity) (Strube&Hahn 1999).

In computational approaches, the linguistic factors are operationalised by means of detectable features which are then used to model (or approximate) local coherence. Some of the underlying concepts overlap to a certain extent; e.g. lexical chains and familiarity, or topichood and syntactic functions are not independent from each other. In order to highlight the impact (and inter-dependence) of the individual features, we will present a comparative evaluation on features that model local coherence, with a special focus on German data. Other comparative evaluations have been carried out in recent work (Barzilay&Lapata 2008, among others). However, most of them focus on English data and take only a subset of features into account.

References

- Barzilay, Regina & Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1). 1-34.
- Grosz, Barbara, Aravind Joshi & Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2). 203-225.
- Strube, Michael & Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics* 25(3). 309-344.

DGfS-CL

Stefan Müller und Felix Bildhauer
Freie Universität Berlin

Informationsstruktur in einer HPSG Grammatik

Aufbauend auf einer umfassenden HPSG-Grammatik für das Deutsche (Müller 2002; 2007) präsentieren wir ein erweitertes Fragment, das auch Aspekte von Informationsstruktur abdeckt. Dazu werden informationsstrukturelle und prosodische Eigenschaften als Merkmal-Wert-Paare modelliert (in Anlehnung an Bildhauer 2008) und Schnittstellenbeschränkungen zwischen Informationsstruktur, Syntax und Phonologie formuliert. Das Fragment ist im TRALE-System (Meurers/Penn/Richter 2002, Penn 2004) implementiert und wird im Rahmen einer Demo vorgestellt.

Literatur:

Bildhauer, Felix. 2008. *Representing Information Structure in an HPSG Grammar of Spanish*. Diss., Universität Bremen.

Meurers, Walt Detmar, Penn, Gerald and Richter, Frank. 2002. A Web-Based Instructional Platform for Constraint-Based Grammar Formalisms and Parsing. In Dragomir Radev and Chris Brew (eds.), *Effective Tools and Methodologies for Teaching NLP and CL*, Proceedings of the Workshop held at 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, Seiten 18-25, http://www.sfs.uni-tuebingen.de/~dm/papers/acl02.pdf, 6.11.2008

Müller, Stefan. 2002. *Complex Predicates: Verbal Complexes, Resultative Constructions, and Particle Verbs in German*. Stanford: CSLI, http://hpsg.fu-berlin.de/~stefan/Pub/complex.html, 6.11.2008.

Müller, Stefan. 2007. *Head-Driven Phrase Structure Grammar. Eine Einführung*. Tübingen: Stauffenburg, http://hpsg.fu-berlin.de/~stefan/Pub/hpsg-lehrbuch.html, 6.11.2008.

Penn, Gerald. 2004. Balancing Clarity and Efficiency in Typed Feature Logic Through Delaying. In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Barcelona, Seiten 239–246, http://www.cs.toronto.edu/~gpenn/papers/when.pdf, 6.11.2008.

DGfS-CL

Eugenie Giesbrecht[1], Stefan Evert[2]
[1]Forschungszentrum Informatik an der Universität Karlsruhe
[2]Institut für Kognitionswissenschaft, Universität Osnabrück

An Evaluation of Part-of-Speech Taggers for the Web as Corpus

Part-of-speech (POS) tagging is an important and widely-used preprocessing step in corpus linguistics and most natural language processing applications. Many computational linguists consider tagging to be a "solved task", with state-of-the-art taggers achieving accuracies around 97% (Schmid, 1995; Toutanova et al., 2003). While this means that, on average, every other sentence contains a tagging mistake, the accuracy is close to the agreement between human annotators and is sufficient for most applications.

These taggers have been trained and evaluated on newspaper text, though, and it is not clear how well they perform on other genres such as literature, spoken language, or Web pages. The latter form a particularly important category, as an increasing number of researchers turn to the World Wide Web as a convenient and inexhaustible source of natural language data (this is often called the "Web as Corpus" approach, see. e.g. Kilgarriff and Grefenstette, 2003).

The goal of the study reported here was to find out whether POS taggers trained on newspaper corpora would perform equally well on Web texts. As there is currently no Web corpus with accurate, manually annotated POS information, we semi-automatically annotated a sample of the Geramn deWaC Web corpus (Baroni and Kilgarriff, 2006). The sample was chosen to contain a representative selection of different genres of Web texts. The TreeTagger (Schmid, 1995), the StanfordTagger (Toutanova et al., 2003) and the Apache UIMA Tagger[1] were trained on the Tiger Treebank (Brants et al., 2002) and evaluated on the deWaC samples.

We did also the text type evaluation. The experiments showed that most texts or genres in a web corpus could be tagged rather well, while some others, specific for the web, were disastrous.

Further, we show that using a simplified tagset results in significantly higher tagging accuracy than using a more fine-grained tagset, especially for the problematic web genres.

References:

Baroni, Marco and Kilgarriff, Adam. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of European Chapter of the ACL (EACL) Conference*.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT - NAACL*.

Kilgarriff and Gregory Grefenstette. 2003. Introduction to the Special Issue on Web as Corpus. *Computational Linguistics,* 29 (3).

[1] http://incubator.apache.org/uima/sandbox.html#tagger.annotator

DGfS-CL

Manu Raster, M.A.
Universität Konstanz

Stammbildungswechsel germanischer Verben

Das germanische Verbalsystem unterscheidet bekanntlich starke und schwache Stammbildung, je nach dem auf welche Weise das Präteritum und das Partizip des Präteritums gebildet werden. Starke Verben zeigen Ablaut und zusätzlich die Verwendung des *-onó- Suffixes beim Partizip Präteritum. Schwache Verben lauten im Allgemeinen nicht ab und bilden die Vergangenheitsformen mittels *-d-.

Für den Englischen Zweig der germanischen Sprachen hat vor kurzem eine nicht-linguistische Forschergruppe Änderungen in der Stammbildung untersucht und dafür Belegmaterial aus einer Zeitspanne von etwa 1200 Jahren ausgezählt[2]: Von 177 starken Verben im Altenglischen (in [1] grob auf 800 n. Chr. festgesetzt) werden bis zum Beginn Mittelenglischer Zeit (ab ca. 1200 n. Chr.) 32 Verben nicht mehr stark, sondern nun schwach flektiert. Weitere 47 wechseln beim Übergang vom Mittelenglischen zum heutigen Englisch von starker zu schwacher Flexion.

Durch Extrapolation der Daten auf der t-Achse lassen sich sowohl Prognosen für die Zukunft, als auch Hypothesen für die Vorgeschichte des Englischen aufstellen. Z.B. lautet eine plausible Vorhersage, dass *wed, wed, wed* das nächste Verb sein wird, welches seine nur noch sehr selten gebrauchten starken Vergangenheitsformen zu Gunsten schwach *wed, wedded, wedded* aufgeben wird.

Aufbauend auf diese bestechend klare Untersuchung für das Englische analysieren wir derzeit die entsprechenden Verhältnisse in weiteren germanischen Sprachen. Die Ergebnisse der laufenden Untersuchung wollen wir auf der DGfS-Tagung in Osnabrück präsentieren.

[2] Lieberman, Erez, Joe Jackson, Jean-Baptiste Michel, Tina Tang, and Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449 713— 716.

Eine Datenbank als *multi-engine*

für Sammlung, Vergleich und Berechnung möglichst verlässlicher unterspezifizierter syntaktisch/semantischer Satzrepräsentationen

Kurt Eberle, Kerstin Eckart, Ulrich Heid
Universität Stuttgart-IMS-CL, SFB-732
Azenbergstraße 12,
70174 Stuttgart
{eberle, eckartkn, heid@ims.uni-stuttgart.de}

Für das Korpus-basierte Studium von semantischen Phänomenen ist es oft sinnvoll, semantische Analysen von Sätzen zur Verfügung zu haben. Unterspezifizierte Repräsentationen sind dabei in der Regel besonders geeignet und verlässlich, weil sie Disambiguierung vermeiden, wo dies für die Untersuchung des betrachteten Phänomens unnötig ist (vgl. (Eberle et al. 2008) und die dortige Untersuchung sortaler Mehrdeutigkeiten von deutschen –ung-Nominalisierungen auf der Basis von (Ehrich, Rapp 2001)). Trotzdem können auch unterspezifizierte Repräsentationen Fehler enthalten, die die quantitative Phänomenanalyse negativ beeinflussen.

Unterschiedlich konzipierte Analysewerkzeuge zeigen oft unterschiedliches Verhalten bei der Analyse von Sätzen, mit unterschiedlichen Schwachstellen. Ein Vergleich der Ergebnisse kann die Schwachstellen und Fehler häufig lokalisieren und erlaubt eine Neuberechnung aus der Zusammenschau, die in den meisten Fällen besser ist als die Einzelergebnisse.
Solche *multi-engine*-Architekturen sind u.a. vorgeschlagen worden zur Qualitätsoptimierung bei der Maschinellen Übersetzung (vgl. Wahlster et al. 2000, Chen et al. 2007).

Wir beschreiben in dem Poster eine Datenbank-Architektur, die erlaubt, die Ergebnisse unterschiedlicher Analysewerkzeuge in verschiedenenen Granularitätsstufen und Sichten zu verwalten, durch eine Hierarchie von Beschreibungsmitteln zu vergleichen, Gemeinsamkeiten und Unterschiede zu erkennen und daraus vereinheitlichte Repräsentationen zu berechnen.
Die Datenbank ist dabei Ressource zur vergleichenden Inspektion durch ein Front-End-Werkzeug und Zielmedium für den Import von Analysen aus der Datenbankoberfläche. Bisher verwendete unterspezifizierte Daten stammen aus der Forschungsversion des Lingenio-*translate*-Systems (Eberle et al. 2008) und des von (Schiehlen 2003) entwickelten Werkzeugs.

Wir demonstrieren in dem Poster diverse Suchanfragen und zeigen, in welcher Weise der Ansatz kompatibel ist mit laufenden Datenbankvorhaben (wie ANNIS vgl. (Chiarcos et al. 2008)) und Formatstandardisierungen (wie PAULA vgl. (Dipper 2005)) und welche gegenseitigen synergetischen Effekte generiert werden können.

Literatur

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus and Silke Theison. 2007. Multi-Engine Machine Translation with an Open-Source SMT Decoder, ACL Workshop on Statistical Machine Translation

C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz and M. Stede . 2008. A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. Iin TAL (Traitement Automatique des Langues), Volume 49 (2)

S. Dipper. 2005. XML-based stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In Proceedings of Berliner XML Tage

Kurt Eberle, Ulrich Heid, Manuel Kountz and Kerstin Eckart. 2008 A Tool for Corpus Analysis using Partial Disambiguation and Bootstrapping of the Lexicon. In: Storrer et al. (eds.): Text Resources and Lexical Knowledge, Mouton de Gruyter, Berlin.

Veronika Ehrich and Irene Rapp. 2000. Sortale Bedeutung und Argumentstruktur: *-ung*-Nominalisierungen im Deutschen, Zeitschrift für Sprachwissenschaft 19(2), S. 245-303

Michael Schiehlen. 2003. A Cascaded Finite-State Parser for German, Proceedings of EACL

Wolfgang Wahlster (ed). 2000.Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin

DGfS-CL

Mareile Knees
(Institut für Germanistische Sprachwissenschaft, Friedrich-Schiller-Universität Jena)

Zur maschinellen Auflösung anaphorischer Bezüge auf abstrakte Entitäten am Beispiel von *danach*

Insgesamt haben sich in der computerlinguistischen Forschung nur wenig Studien (s. Eckert/ Strube 2001, Byron 2002, 2004 Artstein/ Poesio 2006) mit der Auflösung von Anaphern beschäftigt, die auf abstrakte Entitäten (wie Situationen oder Sachverhalten) verweisen. Abstrakte Entitäten werden in der Regel durch satzwertige Ausdrücke beschrieben und sind sowohl sprachlich als auch mental schwerer fassbar (s. (1)) als konkrete Entitäten (wie Personen oder Dinge), die durch NPs beschrieben werden (s. (2)).

(1) *Wir schauten* Top Gun, *diesen Tom-Cruise-Film über amerikanische Kampfpiloten.* ***Danach*** *flogen meine Freunde mit ihren hölzernen Segelgleitern steile Kurven, sie wollten sein wie Tom Cruise.* (ZEIT Campus, 05/2007)

(2) *Zu Obamas Beratern zählt außerdem der legendäre Investment-Milliardär Warren Buffet. Mit 78 Jahren dürfte* ***er*** *aber wohl kein Interesse mehr an einem Wechsel in die Politik haben*. (ZEIT online, 29.10.2008)

Nach aktuellem Stand gelingt es den erfolgreichsten Anaphernauflösungsalgorithmen (z.B. Stuckardt 2001, Hinrichs/ Filippova/ Wunsch 2006) basierend auf rein grammatischen Merkmalen (wie Genus und Numerus) 70-85% aller Anaphern mit NP-Antezedenten in einem Text automatisch aufzulösen. Für Anaphern mit satzwertigen Antezedenten und abstrakten Referenten liegt die Auflösungsquote bei 60-70% (s. Eckert/ Strube 2001 und Byron 2002, 2004). Für die Interpretation der restlichen Anaphern ist die Integration semantischer und konzeptueller Merkmale erforderlich.
Der hier vorgestellte Ansatz zur automatischen Auflösung von *danach*-Bezügen fungiert als exemplarische Untersuchung für anaphorische Bezüge auf abstrakte Entitäten. *Danach* bezieht sich nämlich häufig auf satzwertige Ausdrücke, deren Referent eine abstrakte Entität ist (s. Knees 2008). Es wird gezeigt, wie basierend auf rein grammatischen und strukturellen Beschränkungen (wie textuelle Distanz) die Hälfte aller anaphorischen *danach*-Bezüge automatisch auflösen werden können. Um alle anaphorischen Bezüge aufzulösen, müssen semantische und konzeptuelle Merkmale berücksichtigt werden. Diese sind jedoch nur zum Teil implementierbar.

Artstein, R./ Poesio, M. (2006). Identifying reference to abstract objects in dialogue. In *Brandial 2006 Proceedings*. Potsdam, Germany, September 2006.

Byron, D.K. (2002). Resolving Pronominal Reference to Abstract Entities. *Proceedings of ACL 2002*, 80-87.

Byron, D.K. (2004). *Resolving pronominal reference to abstract entities*. Dissertation, Technical Report 815, University of Rochester, Dept. of Computer Science, January 2004.

Eckert, M./ Strube, M. (2001). Dialogue Acts, Synchronizing Units and Anaphora Resolution. *Journal of Semantics 15*, 51-89.

Hinrichs, E. W./ Filippova, K./ Wunsch, H. (2006). A Data-driven Approach to Pronominal Anaphora Resolution in German. In Nicolov, N./ Bontcheva, K./ Angelova, G./ Mitkov, R. (Hg.), *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP'2005*. Amsterdam/ Philadelphia: Benjamins.

Knees, M. (2008): *Zur Semantik und Referenz des temporalanaphorischen Pronominaladverbs „*danach*“*. Dissertation, Philosophische Fakultät, Universität Jena.

Stuckardt, R. (2001). Design and Enhanced Evaluation of Robust Anaphor Resolution Algorithm. Computational Linguistics 27(4), 479-506.