Linguistics in Computational Linguistics: Observations and Predictions

Hans Uszkoreit
Language Technology Lab, DFKI GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbruecken
uszkoreit@dfki.de

Abstract

As my title suggests, this position paper focuses on the relevance of linguistics in NLP instead of asking the inverse question. Although the question about the role of computational linguistics in the study of language may theoretically be much more interesting than the selected topic, I feel that my choice is more appropriate for the purpose and context of this workshop.

This position paper starts with some retrospective observations clarifying my view on the ambivalent and multi-facetted relationship between linguistics and computational linguistics as it has evolved from both applied and theoretical research on language processing. In four brief points I will then strongly advocate a strengthened relationship from which both sides benefit.

First, I will observe that recent developments in both deep linguistic processing and statistical NLP suggest a certain plausible division of labor between the two paradigms.

Second, I want to propose a systematic approach to research on hybrid systems which determines optimal combinations of the paradigms and continuously monitors the division of labor as both paradigm progress. Concrete examples illustrating the proposal are taken from our own research.

Third, I will argue that a central vision of computational linguistics is still alive, the dream of a formalized reusable linguistic knowledge source embodying the core competence of a language that can be utilized for wide range of applications.

1 Introduction

Computational linguistics did not organically grow out of linguistics as a new branch of mathematical or applied linguistics. Although the term suggests the association with linguistics, in practice much of CL has rather been purely engineering-driven natural language processing. Even if computational linguistics has become a recognized subfield of linguistics, most of the action in CL does not address linguistic research questions.

For most practitioners, the term was never more than a sexy sounding synonym for natural language processing. Many others, however, fortunately including many of the most creative and successful scientists in CL, shared the ambition of contributing to the scientific study of human language.

Already in the eighties Lauri Karttunen observed that there is a coexistence and mutual fertilization of applied computational linguistics and theoretical computational linguistics, and that the latter subarea can provide important insights into the structure and use of human language.

When we look into the actual relationship between linguistics and CL, we can easily perceive a number of changes that have happened over time. We can distinguish five major paradigms in computational linguistics, each of which has assigned a slightly different role to linguistic research. The first paradigm was the direct procedural implementation of language processing. NLP systems of this paradigm were programs in languages such as FORTRAN, COBOL or assembler in which there was no systematic division between linguistic knowledge and processing. Linguistics was only important because it had educated some of the practitioners on relevant properties of human language.

The second paradigm was the development of specialized algorithms and methods for language

Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics, pages 22–25,
Athens, Greece, 30 March, 2009. ©2009 Association for Computational Linguistics

processing. This paradigm includes for instance parsing algorithms, finite-state parsers, ATNs, RTNs and augmented phrase-structure grammars. Although we find a separation between linguistic knowledge and processing components, none of the developed methods were imports from linguistics, nor were they adopted in linguistics. (A notable exception may have been two-level finite-state morphology which at least caused some discussion in linguistic morphology.) Nevertheless, some of the approaches required a certain level of linguistic sophistication.

The third paradigm was the emergence of linguistic formalisms. In the eighties a variety of new declarative grammatical formalisms such as HPSG, LFG, CCG, CUG had quite some influence on CL. These formal grammar models were accompanied by semantic formalisms such as DRT. A number of these formal models were tightly connected to linguistic theories and therefore also taught in linguistics curricula. Several attempts to turn current versions of the linguistic mainstream theory of GB/P&P/minimalism into such a declarative formalism were not very successful in NLP but still discussed and used in linguistic classrooms.

When these linguistic formalisms failed to meet the performance criteria needed for realistic applications, most of applied computational linguistics research fell back on specialized methods for NLP such as finite state methods for information extraction. Other colleagues moved on to methods of the fourth paradigm in CL, i.e., statistical methods. Inspired by the rapid success of these statistical techniques, the new paradigm soon ruled most of NLP research. Not surprisingly, the distance between linguistics and mainstream CL increased, as researchers in most subareas researchers did not have to know much about language and linguistics in order to be successful in statistical NLP.

Only when the success curve of statistical NLP started to flatten in several application areas, interest in linguistic methods and knowledge sources reawakened. Hard core statistical NLP specialists consulted lexicons or tried to develop statistical models on phrase structures. Many statistical approaches now exploit structured linguistic descriptions as obtained from treebanks and other linguistically annotated corpora.

In the meantime, proponents of linguistic methods had discovered the power of statistical models for overcoming some of the performance limitations of deep NLP. Statistical models trained on treebanks have become the preferred method for solving the massive ambiguity problem of deep linguistic parsing.

All these pragmatic mixes of statistical and linguistic methods marked the birth of the fifth paradigm in CL, the creative combination of statistical and non-statistical machine learning approaches with linguistic methods.

2 Division of Labor between Linguistics and Statistics

To illustrate my view on the complementary contributions of statistical and linguistic methods I want to start with three observations. The first observation stems from parser evaluations. A CCG parser was successfully applied to the standard Wall Street Journal test data within the Penn Treebank (refs). Although the C&C parser did not quite get the same coverage as the best statistical systems, it produced very impressive results. As Mark Steedman demonstrated in a talk at the Computational Linguistics session of the 2008 International Congress of Linguists, the C&C parser moreover found many dependencies needed for semantic interpretation that are not even annotated in the Penn Treebank.

Observation two stems from our work on hybrid machine translation. Within the EU project EuroMatrix we are organizing open evaluation campaigns of MT systems by shared tasks whose results are reported in the annual WMT workshops. The first large campaign combining automatic and intellectual evaluation took place in 2008. Participants could contribute translations of two test data sets for a range of language pairs. One test set was in a specific domain for which training data had been provided. The other test set contained news texts on a variety of topics. Although a training set of news texts had been provided as well, the covered domains exhibited much more diversity than the closed domain texts. It turned out that in general the best systems for the closed domain task were statistical MT systems, whereas the open domain task was best solved by seasoned rule-based MT commercial products.

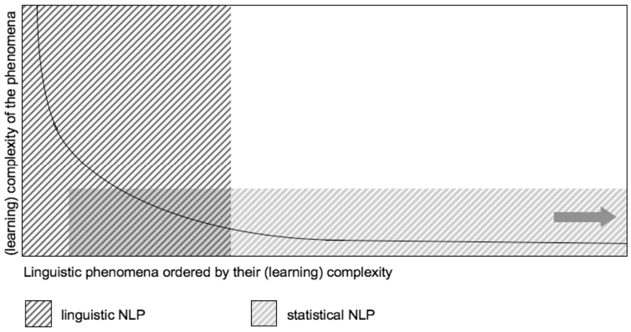
A careful comparative study of errors made by some of the best SMT and RBMT systems revealed that the errors of the two systems were largely complementary. As SMT can acquire frequently used expressions from training data, the output generally appears rather fluent, at least for short sentences and short portions of sentences. SMT is also superior in lexical and phrasal disambiguation and the optimal lexical choice in the target language. However, the translations exhibit many syntactic problems such as missing verbs or agreement violations, especially if the target language has a complex morphology. RBMT systems, on the other hand, usually get the syntactic structure right–unless they fail in attachment ambiguities–but on the word and phrase level they often do not select the correct or stylistically optimal translations.

Today's machine learning methods for acquiring the statistical translation models from parallel texts fail on many syntactic phenomena that can be analyzed correctly by a linguistic grammar. Inducing a correct treatment of long distance phenomena such as topicalization or "easy"-adjectives, ellipsis and control phenomena from unannotated texts seems quite impossible. Learning complex rules from syntactically and semantically annoted texts may be possible if linguists have already understood and formalized the underlying analysis of the phenomenon.

The third observation comes from supervising work in grammar development and attempts to enlarge the coverage of existing grammars automatically through the exploitation of corpus data. When he tried to extend the coverage of the ERG, Zhang Yi could show that almost all of the coverage gaps could be attributed to missing lexical knowledge. Even if the words in the unanalyzable sentence were all in the lexicon, usually some reading of words, i.e. their membership in some additional word class, was missing. The few remaining coverage deficits result from specific infrequent constructions not yet covered by the grammar plus missing treatments for a few notoriously tricky syntactic problems such as certain types of ellipsis.

These three observations together with numerous others strongly suggest the following insight. Every grammar of a human language consists of a small set of highly complex regularities and of a huge set of much less complex phenomena. The small set of highly complex phenomena occurs much more often than most of the phenomena of little complexity. This slanted distribution makes language learnable. So far we have no automatic learning methods that could correctly induce the complex phenomena. It is highly questionable whether these regularities could ever be induced without full access to the syntax-semantics mappings that the human language learner exploits.

On the other hand, the lexicon or simple selectional restrictions can easily be learned because the complexity lies in the structure of the lexical classes and not in the simple mapping from words to these classes.



3 Hybrid Systems Research

In several areas of language processing, first approaches of designing hybrid systems containing both linguistic and statistical components have demonstrated promising results.

However, much of this research is based on rather opportunistic selections. Readily available components are connected in a pure trial and error fashion. In our hybrid MT research we are systematically searching for optimal combinations of the best statistical and the best rule-based systems for a given language pair. The approach is systematic, because we use a detailed error analysis by skilled linguists to find out which classes of phrases are usually better translated by the best statistical systems. We then insert the translations for such kinds of phrases into the syntactic skeletons of the translated sentences provided by the rule-base system. One of the translations we submitted to this year’s EuroMatrix evaluation campaign was obtained in this way.

The technique to merge sentence parts from the two systems into one translation is only a crude first approximation of a truly hybrid processing system, i.e., a system in which the statistical phrase translation is fully integrated into the rule-based system. Our goal is to test the usefulness of statistical methods in analysis, especially for disambiguation, in transfer, especially for selecting the best translation for words and smaller phrases and in generation, for the selection among paraphrases according to monolingual language models.

Another systematic approach to hybrid systems design was investigated in the Norwegian LOGON project, in which deep linguistic processing by HPSG and LFG was complemented by statistical methods.

Another example for a systematic approach to hybrid systems building is our work on an architecture for the combination of components for the analysis of texts. The DFKI platform Heart-of-Gold (HoG) was especially designed for this purpose. In HoG several components can be combined in multiple ways. All processing components write their analysis results into a multi-layer XML standoff annotation of the analyzed text. The actual interface language is RMRS (Robust Minimal Recursion Semantics, ref.) XML is just used as the syntactic carrier language for RMRS.

4 Computational Models of Linguistic Competence

Although the competence-performance distinction is a complex and highly controversial issue, the theoretical dichotomy is useful for the argument I want to make. When children acquire a language, they first learn to comprehend and produce spoken utterances. Much later they learn to read and to write, and much later again they may learn how to sing and rhyme and how to summarize, translate and proofread texts.

All of the acquired types of performance utilize their underlying linguistic competence. New types of performance are relatively easy to learn. The shared knowledge base ensures a useful level of consistency across the performance skills. Of course, each type of performance may use different parts of the shared competence. Certain types of performance may also extend the shared base into different directions.

The child could not acquire the complex mapping between sound and meaning without having access to both spoken (and later also written) form and the corresponding semantics. Therefore the child cannot learn a language from a radio beside her crib, nor can the older child acquire Chinese by being locked up in a library of Chinese books. Thus the basic competence cannot be obtained outside performance or successful communication.

The first approaches to linguistic computational grammars may have been too simplistic by not providing the connection between competence and performance needed for exploiting the competence base in realistic applications. However, in gradually solving the problems of efficiency, robustness and coverage researchers have arrived at more sophisticated views of deep linguistic processing.

After several decades of experience in working on competence and performance modeling for both generic grammatical resources and many specialized applications, I am fully convinced that the goal of a reusable shared competence model for every surviving language in our global digital information and communication structure is still a worthwhile and central goal of computational linguistics. I am also certain that the goal will be obtained in many steps. We already witness a reuse of large computational grammar resources such as the HPSG ERG, the LFG ParGram Grammar and the English CCG in many different applications. These applications are still experimental but when deep linguistic processing keeps improving in efficiency, specificity (ability to select among readings), robustness and coverage at current speed of progress, we will soon see first cases of real life applications.

I am not able to predict the respective proportions of the intellectually designed core components, the components learned automatically from linguistically annotated data and the components automatically learned from unannotated data but I am convinced that the systematic search for the best combinations will be central to partially realizing the dream of computational linguistics still within our life times.

If such solutions can be found and gradually improved, the insights gained through this systematic investigation may certainly also have a strong impact in the other direction, i.e. from computational linguistics into linguistics.