Towards a Road Map on Human Language Technology: Natural Language Processing

Editors: Andreas Eisele, Dorothea Ziegler-Eisele

Version 2 (March 2002)

Abstract

This document summarizes contributions and discussions from two workshops that took place in November 2000 and July 2001. It presents some visions of NLP-related applications that may become reality within ten years from now. It investigates the technological requirements that must be met in order to make these visions realistic and sketches milestones that may help to measure our progress towards these goals.

1. Introduction

Scope of this Document

One of the items on ELSNET's agenda for the period 2000-2002 is to develop views on and visions of the longer-term future of the field of language and speech technologies and neighboring areas, also called ELSNET's Road Map for Human Language Technologies. As a first step in this process, ELSNET's Research Task group is organizing a series of brainstorming workshop with a number of prominent researchers and developers from our community. The first one of these workshops took place in November 2000 under the general motto “How will language and speech technology be used in the information world of 2010? Research challenges and infrastructure needs for the next ten years". The second one was co-organized in July 2001 by ELSNET and MITRE as part of ACL-2001 and had the somewhat more specific orientation on “Human Language Technology and Knowledge Management (HLT-KM)”. This workshop brought together more than 40 researchers from industry and academia and covered a considerable range of topics related to KM and HLT in general.

This paper aims at summarizing and organizing material from both workshops, but concentrates on applications and technologies that involve NLP, i.e. the processing of written natural language, as speech-related technologies and new models of interactivity have already been covered in documents presented around the first workshop. In the discussion of question answering and summarization, vision papers and roadmaps compiled by researchers in the US and published by NIST have been taken as an additional source of inspiration.

The Growing Need for Human Language Technology

Natural language is the prime vehicle in which information is encoded, by which it is accessed and through which it is disseminated. With the explosion in the quantity of on-line

text and multimedia information in recent years there is a pressing demand for technologies that facilitate the access to and exploitation of the knowledge contained in these documents.

Advances in human language technology will offer nearly universal access to on-line information and services for more and more people, with or without skills to use computers. These technologies will play a key role in the age of information and are cited as key capabilities for competitive advantage in global enterprises.

Extraction of knowledge from multiple sources and languages (books, periodicals, newscasts, satellite images, etc.) and the fusion into a single, coherent textual representation requires not only an understanding of the informational content of each of these documents, the removal of redundancies and resolution of contradictions. Also, models of the user are required, the prior knowledge that can be assumed, the level of abstraction and the style that is appropriate to produce output that is suitable for a given purpose.

More advanced knowledge management (KM) applications will be able to draw inferences and to present the conclusions to the user in condensed form, but let the user ask for explanations of the internal reasoning. In order to find solutions for problems beyond a static pool of knowledge, we need systems that are able to identify experts, who have solved similar problems. Again, advanced NLP capabilities will be required to appraise the aptitude of candidates from documents authored by them or describing prior performance.

But also outside of KM, sophisticated applications of NLP will emerge over the next years and decades and find their way into our daily lives. The range of possibilities is almost unlimited. An important group of applications is related to electronic commerce, i.e. new methods to establish and maintain contact between companies and their customers. Via mobile phones, e-mail, animated web-based interfaces, or innovative multi-channel interfaces, people will want to make use of all kinds of services related to buying and selling goods, home-banking, booking of journeys, and the like. Also in the area of electronic learning a considerable growth is expected within the coming years.

Multilinguality

Whereas English is still the predominant language on the WWW, the fraction of non-English Web pages and sites is steadily increasing. Contrasting earlier apprehensions, the future will probably present ample opportunities for giving value to different languages and cultures. However, the possibility to collect information from disparate, multilingual sources also provides considerable challenges for the human user of these sources and for any kind of NLP technology that will be employed.

One of the major challenges is lexical complexity. There will be about 200 different languages on the web and thus about 40.000 potential language pairs for translation. Clearly, it will not be possible to build bilingual dictionaries that are comprehensive both in the number of language pairs and in the coverage of application domains. Instead, multilingual vocabularies need to provide mappings into language independent knowledge organization structures, i.e. common systems of concepts linked by semantic relations. However, the definition of such an “interlingua” will be difficult in cases in which languages make distinctions of different granularity.

Research Trends and Challenges

The field of human language technology covers a broad range of activities with the goal of enabling people to communicate with machines using natural communication skills.

Although NLP can help to facilitate knowledge management, it requires a large amount of specialized knowledge by itself. This knowledge may be encoded in complex systems of linguistic rules and descriptions, such as grammars and lexicons, which are written in dedicated grammar formalisms and typically require many person-years of development effort. The rules and entries in such descriptions interact in complex ways, and adaptation of such a sophisticated system to a new text style or application domain is a task that requires a considerable amount of specialized manpower.

One way to cope with the difficulties in the acquisition of linguistic knowledge was to restrict attention to shallower tasks, such as looking for syntactic "chunks" instead of a full syntactic analysis. Whereas this has proven rather successful for some applications, it obviously severely limits the depth to which the meaning of a document or utterance is taken into account.

Another approach was to shift attention towards models of linguistic performance (what occurs in practice, instead of what is principally possible) and to use statistical or machine learning methods to acquire the necessary parameters from corpora of annotated examples. These data-driven approaches offer the possibility to express and exploit gradual distinctions, which is quite important in practice. They are not only easier to scale and adapt to new domains, their algorithms are also inherently robust, i.e. they can deal, to a certain extent, gracefully with errors in the input.

Statistical parsers, trained on suitable tree banks, now achieve more than 90% precision and recall in the recognition of syntactic constituents in unseen sentences from English financial newspaper text.

However, a lot of work remains to be done, and it is not obvious how the success of corpus-driven approaches can be enlarged along many dimensions simultaneously. One challenge is that analysis methods need to work for many languages, application domains and text types, whereas the manual annotation of large corpora of all relevant types will not be economically feasible. Another challenge is that, other than syntax, many additional levels of analysis will be required, such as the identification of word sense, the reference of expressions, structure of argumentation and of documents, and the pragmatic role of utterances. Often, the theoretical foundation that is required before the annotation of corpora can begin is still lacking.

One could say that for corpus-driven approaches the issue of scalability of the required resources shows up again, albeit in a somewhat different disguise. Hence, research in NLP will have to address this issue seriously, and find answers to the question how better tools and learning methods can reduce the effort of manual annotation, how annotated corpora of a slightly different type could best be re-used, how data-driven acquisition processes can exploit and extend existing lexicons and grammars, and finally how analysis levels for which the theoretical basis is still under development could be advanced in a data-driven way.

Structure of this Document

The remainder of this document is structured as follows. In Chapter 2 we describe a number of prototypical applications and scenarios in which NLP will play a crucial role. Whereas each of these scenarios is discussed mainly from a user's perspective, we also give indications, which technological requirements must be met to make various levels of sophistication of these applications possible. In Chapter 3, the technologies that have been mentioned earlier are discussed in more detail, and we try to indicate which levels of functionality may be expected within the timeframe of this study. These building blocks are

then put into a tentative chronological order, which is displayed in Chapter 4. Finally, Chapter 5 gives some general recommendations about beneficial measures concerning the infrastructure for the relevant research.

2. Applications of NLP

Recent developments in natural language processing have made it clear that formerly independent technologies can be harnessed together to an increasing degree in order to form sophisticated and powerful information delivery vehicles. Information retrieval engines, text summarizers, question answering and other dialog systems, and language translators provide complementary functionalities which can be combined to serve a variety of users, ranging from the casual user asking questions of the web to a sophisticated, professional knowledge worker.

Though one cannot strictly separate the following applications from each other, because one can act as a part of another, we try to dissect the large field of existing and future applications in the hope of making the field as a whole more transparent.

Information Retrieval (IR)

What is called information retrieval today is actually but a foretaste of what it should be. Current systems neither understand the information need of the user, nor the content of the documents in their repositories. Instead of meaningful replies, they just return a ranked, and often very long list of documents that are somehow related to the given query, which is typically very short. A better name for this restricted functionality would be text retrieval.

Information retrieval systems must understand a query, retrieve relevant information, and present the results. Retrieved information may consist of a long document, multiple documents of the same topic, etc and good systems should present the most important material in a clear and coherent manner.

Current information retrieval techniques either rely on an encoding process using a certain perspective or classification scheme to describe a given item, or perform a superficial full-text analysis, searching for user-specific words. Neither case guarantees content matching.

The ability to leverage advances in input processing (especially natural language query processing) together with advances in content-based access to multimedia artifacts (e.g., text, audio, imagery, video) promises to enhance the richness and breadth of accessible material while at the same time improving retrieval precision and recall and thus reducing the search time. Dealing with noisy, large scale, and multimedia data from sources as diverse as radio, television, documents, web pages, and human conversations (e.g., chat sessions and speech transcriptions) will offer challenges.

One important part of IR would be multi-document summarization that can turn a large set of input documents into several different short summaries, which can then be sorted by topics or otherwise put into a coherent order.

Summarization

Summarization will enable knowledge workers access to larger amounts of material with less required reading time. The goal of automatic text summarization is to take a partially structured source text, extract information content from it and present the most important content in a condensed form in a manner sensitive to the needs of the user and task. Scalability to large collections and the generation of user-tailored or purpose-tailored summaries are active areas of research.

The summarization can either be an extract consisting entirely of material copied from the input, or an abstract containing material not present in the input, such as subject categories, paraphrases of content, etc.

For extraction shallower approaches are possible, as frequently the sentences may be extracted out of context. The transformation here involves selecting salient units and synthesizing them with the necessary smoothing (adjusting references, rearranging the text…). Training by using large corpora is possible.

Abstracts need a deeper level of analysis, the synthesis involves natural language generation and some coding for a domain is required.

Depending on their function, three types of abstracts can be distinguished: An indicative abstract provides a reference function for selecting documents for more in-depth reading. An informative abstract covers all the salient information in the source at some level of detail and evaluative abstracts express the abstractor's views on the quality of the work of the author.

Characteristics for the summarization are the reduction of the information content (compression rate), the fidelity to the source, the relevance to the user's interest, and the well-formedness regarding both to syntactic and discourse level. Extracts need to avoid gaps, dangling anaphora, ravaged tables and lists, abstracts need to produce grammatical, plausible output.

Some current applications of summarization are:

1. Multimedia news summaries: watch the news and tell what happened while I was away
2. Physicians' aids: summarize and compare the recommended treatments for this patient
3. Meeting summarization: find out what happened at that teleconference I missed
4. Search engine hits: summarize the information in hit lists retrieved by search engines
5. Intelligence gathering: create a 500-word biography of Osama bin Laden
6. Hand-held devices: create a screen-sized summary of a book
7. Aids for the Handicapped: compact the text and read it out for a blind person

Though there are already promising approaches towards mastering all types of summaries, there are still obstacles to overcome such as the need for robust methods for the recognition of semantic relations, speech acts, and rhetorical structure.

Question Answering (QA)

The straightest way to get access to the gigantic volume of knowledge around us is probably asking questions by communicating with other persons, computers or machines.

An important new class of systems will move us from our current form of search on the web (type in keywords to retrieve documents) to a more direct form of asking questions in natural language, which are then directly responded to with an extracted or generated answer. Currently it is rather straightforward to get an answer to "what questions" (what is the capital of China, what are the opening hours of the hermitage etc.), whereas "why questions" (why did the new market fail) are normally not answered by an information retrieval query, unless the answer happens to be present in the information database, or can be inferred afterwards by the user from the answers she gets.

In the next decade time has come to find answers to why questions from information systems by letting the systems make the appropriate inferences. This requires very sophisticated automatic reasoning methods, based on systematic extraction of information from texts, storing the information in a systematized way, which lends itself to reasoning and inference rules that will be able to draw the proper conclusions from the knowledge stored in the information database.

We can subdivide the long-term goal of building powerful, multipurpose information management systems for QA in simpler subtasks that can be attacked in parallel at varying levels of sophistication, over shorter time frames.

Clearly there is not a single, archetypical user of a Q&A system. In fact there is a full spectrum of questions, starting with simple factual questions, which could be answered in a single short phrase found in a single document (e.g. "Where is the Taj Mahal?"). Next, questions like "What do we know about Company xyz?", where the answer cannot be found in a single document but will require retrieving multiple documents, locating portions of answers in them and combining them into a single response. This kind of question might be addressed by decomposing it into a series of single focus questions.

Finally there are very complex questions, with broad scope, using judgment terms and needing deep knowledge of the user's context to be answered. Imagine someone is watching a television newscast, becomes interested in a person, who appears to be acting as an advisor to the country's Prime Minister. And now the person wants to know things like: "Who is this individual. What is his background? What do we know about the political relationship of this person and the Prime Minister and/or the ruling party?". The future systems that can deal with this type of questions must manage the search in multiple sources in multiple media/languages, the fusion of information, resolution of conflicting data, multiple alternatives, adding interpretation, drawing conclusions.

In order to realize this goal, research must deal with question analysis, response discovery and generation from heterogeneous sources, which may include structured and unstructured language data of all media types, multiple languages, multiple styles, formats and also image data i.e. document images, photography and video.

To the extent to which NLP research will learn to master the challenges of source selection, source segmentation, extraction, and semantic integration across heterogeneous sources of unstructured and semi-structured data, NLP technology will help us to reduce the time,

memory, and attention required to sift through many returned web pages from a traditional search by providing direct answers to questions.

Semantic Web

The standardization committee for the WWW (called W3C) expects around a billion web users by 2002 and an even higher number of available documents. However, this success and exponential grow makes it increasingly difficult to find, to access, to present, and to maintain the information of use to a wide variety of users.

The semantic web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.

The semantic web is not a separate web but an extension of the current one, in which information is given well-defined meaning better enabling computers and people to work in cooperation. With the help of ontologies large amounts of text can be semantically annotated and classified.

Currently pages on the web use representations rooted in format languages such as HTML or SGML. The information content, however, is mainly presented by natural language. Thus, there is a wide gap between the information available for tools that try to address the problems above and the information kept in human readable form.

The semantic web will provide intelligent access to heterogeneous and distributed information enabling software agents to mediate between the user needs and the available information sources.

The first steps in weaving the semantic web into the structure of the existing web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and "understand" the data that they merely display at present.

What is required: creation of a machine understandable semantics for some or all of the information presented in the WWW i.e.

- Developing languages for expressing machine understandable meta-information for documents, in the line of RDF, DAML, and similar proposals.
- Developing terminologies (i.e., name spaces or ontologies) using these languages and making them available on the web.
- Integrating and translating different terminologies
- Developing tools that use such languages and terminologies to provide support in finding, accessing, presenting and maintaining information sources.

Developing such languages, ontologies and tools is a wide-ranging problem that touches on the research areas of a broad variety of research communities.

Creation of the relevant tools will require a better knowledge of what the users want to know from websites, i.e. these developments need to be based on a user-centered process view.

Another crucial issue will be: “Who is going to populate the semantic web?” The semantic markup that is required by automated software agents needs to be very easy to create and supporting tools need to be provided, otherwise this wonderful idea will not have significant impact for a long time. Advanced NLP technology that can “guess” the correct semantic annotation and propose suitable markup semi-automatically will enable conformance to the needs of software agents with minimal manual effort.

Dialogue Systems

No matter if people want to buy something, find or use a service or just need information, dialog systems promise user-friendly and effective ways to achieve these goals, even for first time users.

Despite the apparent resemblance to QA systems, there are several specific problems to be solved concerning dialogue modality and structure. Input to a dialog system might be via keypad, voice, pointing device, combinations thereof, or other channels, so all errors and incompleteness of spontaneous natural language will show up. In contrast to QA systems, there will be mixed initiatives of speaker and system and the scope is much wider if we take into account that the focus during natural dialogue may often change. Also, the utterance made during a dialog can only be correctly interpreted based on the dialog context and the mutual knowledge that has been accumulated before it was made.

In future we require systems that can support natural, mixed initiative human computer interaction that deals robustly with context shift, interruptions, feedback and shift of locus or control.

Open research challenges include the ability to tailor flow and control of interactions and facilitate interactions including error detection and correction tailored to individual physical, perceptual and cognitive differences.

Motivational and engaging life-like agents offer promising opportunities for innovation.

Agent/user modeling: Computers can construct models of user beliefs, goals and plans as well as models of users’ individual and collective skills by processing materials such as documents or user interactions/conversations. While raising important privacy issues, modeling users or groups of users unobtrusively from public materials or conversations can enable a range of important knowledge management capabilities

tracking of user characteristic skills and goals enhances interaction as well as discovery of experts by other users or agents

A central problem for the development of dialogue systems is the fact that contemporary linguistics is still struggling to achieve a genuine integration of semantics and pragmatics. A satisfactory analysis of dialogue requires in general both semantic representation i.e. representation of the content of what the different participants are saying and pragmatic information, i.e. what kinds of speech acts they are performing (are they asking a question, making a proposal…)

Analysis of a dialog needs to explain the purpose behind the utterances it consists of. Determining the semantic representation of an utterance and its pragmatic features must in general proceed in tandem. A dialogue system identifying the relevant semantic and pragmatic information will thus have to be based on a theory in which semantics and

pragmatics are both developed with the formal precision that is a prerequisite for implementation and suitably attuned to each other and intertwined.

Applications in Electronic Commerce

New technological possibilities can quickly impact the interaction between companies and their customers. One example are dialog systems that allow customers to obtain personal advises or services. For reasons indicated above, these systems are difficult to build, but once this investment has been done, they can be operated at low cost for the company.

Another example, which may be even sooner to come, is the creation of systems that support processing of emails sent by customers. According to business analyses, e-mail has already now become one of the most common forms of customer communication. For numerous businesses that are not well-prepared, this has transformed e-mail into a severe pain point, giving rise to the pressing need to adopt e-mail response management systems.

Obviously, NLP technologies that are able to extract the salient facts from email messages can constitute a central part of these systems. Due to the potential complexity of the queries and additional problems like ungrammatical input and spelling errors, the correct interpretation of arbitrary messages is far from easy. However, there are several factors that alleviate the situation: Messages that are too difficult for automatic processing can be routed to human agents. In cases in which doubts about the correctness of generated responses persist, these responses can always be checked by manual inspection. Historical data about email exchange with customers can be used to bootstrap the models that are required for the system. Depending on the business, a significant fraction of the emails may be amenable to NLP, including requests for information material, business reports, certificates, statements of account, scheduling requests, conference registrations etc.

e-Learning

Using modern technology to facilitate learning is one of the most promising application domains of NLP. Good QA systems that are able to give answers to the point, or summarization systems that can adapt to the user's prior knowledge and present important additions in a way that is easy to understand could immediately take the place of a good teacher, which an unlimited supply of time and patience. One technology is ripe to build these tools, using them for e-learning will one of the biggest opportunities to our knowledge society.

However, as the European society evolves more and more into multilingualism, it is natural to ask how NLP can help to make language learning easier and more effective. We can imagine systems to help train children to write and to speak a foreign language. There will be combinations of multi-modal aids for the handicapped. A child will write a sentence and the system will correct it and tutor him about the problems. A child will read a text aloud and the system will monitor which words are not right and why and will analyze where the pronunciation problems are. Later the system would suggest some pronunciation exercises in the particular problem.

Systems that are able to guess the intention of a speaker from the speaker's utterances in a flexible and intelligent way will offer a plethora of possibilities for e-learning. As similar capabilities are required for dialog systems in general, there will be significant synergy effects between these fields of research.

Translation

The idea of machine translation (MT) has been one of the driving forces in the early days of NLP. However, even after more than 50 years of effort, current systems still produce output of limited quality, which is suitable for assimilation of foreign-language documents, but not for the production of publishable material. But even if the old dreams did not come true, MT will play an increasing role in the multilingual world.

Last year, for the first time, English constituted less than half the material on the web. Some predict that Chinese will be the primary language of the web by 2007. Given that information on the web will increasingly appear in foreign languages and not all users will be fluent in those languages, there will be a need to gist or skim content for relevance assessment and/or provide high quality translation for deeper understanding. Some forms of translation for information access is already today available in the web at no cost. The increasing demand for these services will give a push to improve their quality and the providers will find ways to increase vocabularies and translation quality semi-automatically from terminological resources, bilingual corpora and similar sources. Also the need for interactive systems that can give rough translations of chat sessions in real time will create interesting challenges.

Clearly, any systematic collection of lexical and terminological information in the form of domain-specific ontologies will help to build better MT systems for these domains. Conversely, the construction of ontologies can be facilitated by automatic alignment of existing translations, as this will naturally lead to a clustering of the vocabulary along the relevant semantic distinctions.

These developments will also have an impact on improved systems for high-quality translation for the dissemination of documents. Chances are that hybrid combinations of symbolic and stochastic translation engines, able to learn relevant terminology from translation memories will eventually achieve a level of performance that will make them useful for the professional translator. Combined with multi-modal workbenches where voice input, keyboard and mouse interaction will make the composition of the target text as convenient as possible, these new technologies may help at least in some easier domains, where so far the effort of the human translator is dominated by low-level activities such as entering the text, adjusting the formatting, copying names and numbers, which are clearly amenable to partial automation.

3. Technologies for NLP

This chapter contains a more detailed discussion of some of the technologies that are required for the applications mentioned in the last chapter. Most of the material is organized along traditional fields of research in NLP, describing technologies that already exist, but must be further developed to achieve the ambitious goals. Some technologies cannot be assigned to one specific level, because they serve a more generic purpose, such as the extraction of relevant knowledge from text corpora.

Low-Level Processing

Most systems that analyse natural language text typically start by segmenting the text into meaningful tokens. Sometimes, the exact spelling of these tokens needs to be brought into a

canonical form, so that it can match with a lexical entry. Both processes can be based on matching the input against regular expressions, for which efficient algorithms exist. Whereas this task looks straightforward from the distance, there are actually some subtle details that need to be considered. Quite often, a decision whether a word should be split at a special character or whether a dot ends a sentence or is part of the preceding word depends on the vocabulary of the domain and on layout conventions used in this document, so that general rules cannot be defined. Documents that need to be analyzed may contain markup from text processors, which needs to be stripped or interpreted in a suitable way. The knowledge required in these preliminary stages of processing can already be quite specific, so that a manual creation of suitable rule systems is not economically feasible.

Current research on the automatic tokenization and normalization of texts therefore concentrates on the question how the knowledge required by these methods can automatically be derived from examples, using techniques statistical or machine learning approaches.

Another difficulty is the treatment of noise in the input. Output of speech recognition systems often contains recognition errors at rather high rates. Utterances entered interactively or printed documents that have undergone OCR have similar problems. Unfortunately, the distortion of even a single character can mess up the linguistic analysis of the complete input. But of course, we expect NLP systems to deal gracefully and intelligently with small distortions and errors in the input.

To make systems more robust against noisy input, probabilistic techniques for the restoration of distorted signals, which have shown to be quite effective in speech recognition, need to be adapted and generalized to new applications. However, training simple-minded statistical models on massive amounts of data will often not be feasible. By now, statistical language models that incorporate grammatical knowledge are able to give slight improvements over n-gram approaches, and it seems plausible to expect that future improvements of these will be easier to use in specific situation where training data is scarce. Large vocabularies, many types of distortions, and the need to use fine-grained contextual knowledge for improved predictive models constitute significant research challenges. Most likely, there will be some synergy between language models used in speech and similar models that will be developed for low-level processing and correction of written ill-formed input.

Once the segmentation into basic units has been performed, the next step is to identify suitable lexical entries for each token and, in cases where more than one entry applies, to determine which one is most appropriate in the given context. This process is called part-of-speech disambiguation or POS tagging and is usually done with statistical models or machine-learning approaches trained on manually tagged data. Current technology achieves rather high accuracy on newspaper text, but again, performance suffers significantly when a model trained on a certain set of data is applied to text from a different domain. As the output of the POS tagger is typically used as input to subsequent modules, tagging errors may hamper the correct analysis of much more than the affected word. Research on high-quality POS tagging will face problems that are similar to those of language modelling: It requires detailed information about a large number of rare words that may be quite specific to the given domain and application, which is difficult to construct, no matter which road to lexical acquisition is taken. Any effort that will support the construction, distribution, sharing and re-use of large, domain-specific lexical resources will doubtlessly also help to improve the accuracy of POS tagging on text from these domains.

The next step in the analysis of text is to identify groups of words that belong together and refer to one semantic entity. Often, these phrases contain names, and for many practical applications, it is important to classify these expressions according to the type of entity they denote (Person, City, Company, etc.). Depending on the application, the classification may be more or less fine-grained. Again, it is obvious that improved lexical knowledge will help to improve the performance of named entity recognition. But we cannot in all cases rely on a lexical resource to cover the relevant entities. A text may discuss the opening of a new company, which will therefore not be contained in the lexicon. To handle such cases intelligently, we need mechanisms that can exploit contextual clues for the correct classification of unknown entities and we need effective mechanisms that propagate information about new entities into the lexical repositories, so that the system as a whole learns from the texts it sees, similar to the way a human reader would do.

Syntactic Analysis

The goal of syntactic analysis is to break down given textual units, typically sentences, into smaller constituents, to assign categorical labels to them, and to identify the grammatical relations that hold between the various parts.

In most applications of language technology the encoded linguistic knowledge, i.e. the grammar, is separated from the processing components. The grammar consists of a lexicon, and rules that syntactically and semantically combine words and phrases into larger phrases and sentences.

Several language technology products on the market today employ annotated phrase-structure grammars, grammars with several hundreds or thousands of rules describing different phrase types. Each of these rules is annotated by features and sometimes also by expressions in a programming language.

The resulting systems might be sufficiently efficient for some applications but they lack the speed of processing needed for interactive systems, such as applications involving spoken input, or systems that have to process large volumes of texts, as in machine translation.

In current research, a certain polarization has taken place. Very simple grammar models are employed, e.g. different kinds of finite-state grammars that support highly efficient processing. Some approaches do away with grammars altogether and use statistical methods to find basic linguistic patterns. Other than speed, these shallow and statistically trained approaches have advantages in terms of robustness, and they also implicitly perform disambiguation, i.e. when more than one analysis is possible, they make a decision for one reading (which of course may be the wrong one).

On the other end of the scale, we find a variety of powerful linguistically sophisticated representation formalisms that facilitate grammar engineering. These systems are typically set up in a way that all logically possible readings are computed, which increases the clarity (no magic heuristics hidden in procedures), but also slows down the processing. Despite their nice theoretical properties it has so far been difficult to adapt these systems to the needs of real-world applications, where speed, robustness, and partial correctness in typical cases are more urgent than theoretical faithfulness and depth of analysis.

How will this situation evolve? The two approaches will continue to compete for potential applications, and the current advantage for shallow approaches will diminish as more ambitious applications get within reach, and as languages are used that require richer analysis.

This will give incentives for shallow approaches to struggle for higher accuracy and more detailed analyses, whereas the deep processing will be forced to find workable solutions for the problems with speed and robustness. In the ideal case, more fine-grained forms of integration will be found, i.e. hybrid systems that will keep the advantages of both worlds as far as possible.

The simplest integration will just use shallow analysis as a fallback mechanism when deep analysis fails. In this case, results from both approaches need to be translated into one common representation, and the development of such a "common denominator" will be a significant challenge. To achieve an even more fine-grained cooperation between both approaches, deep analysis may be equipped with the ability to locally fall back to more superficial processing, driven by the need to deal with a specific problem in the input. Vice versa, the results of shallow analysis might be combined into a more detailed structure incrementally, based on rules from a deep grammar. Also analyses of corpus data obtained with shallow tools can be mined for linguistic knowledge that is then fed into resources used by a deep parser, and vice versa.

Research challenges will be how to find syntactic parsers that are at the same time fast, robust, deliver a detailed analysis that is correct with high probability and that are easily to adapt to special domains.

Semantic Analysis

The goal of semantic analysis is to assign meanings to utterances, which is an essential precondition for most applications of NLP. However, what level of abstraction is required in this phase depends on the difficulty of the task. Extraction of answers to simple factual questions from a given text will require less depth in analysis than the summarization of a lengthy treatise in few paragraphs.

We can dissect the task of semantic analysis into several subtasks, depending on the linguistic level where it takes place. Most important are the semantic tagging of ambiguous words and phrases, and the resolution of referring expressions.

The disambiguation of word senses needs to identify the meaning that should be assigned to a given word. The hardest part of this task is to define the set of meanings that should be considered in this task, i.e. to select the appropriate granularity for the conceptualization. The emergence of standardized, large-scale ontological resources will help to solve this part of the task, as the concepts that appear in such ontologies are a natural choice for the meanings of single words or simple phrases. Additionally, multilingual corpora that are aligned on the level of words and phrases can serve as an approximation to sense-tagged corpora, so draft ontologies and models for sense disambiguation can be extracted from these.

Considerable efforts in defining useful evaluation metrics for sense disambiguation are pursued in the ongoing SENSEVAL activities. So far, the methods used by the participants of SENSEVAL are mostly based on simple statistical classification using features extracted from the context of word occurrences. To the extent to which robust, high quality systems for syntactic analysis will appear, this will also help to obtain improved accuracy in the semantic disambiguation.

The resolution of referring expression such as pronouns or definite noun phrases is the ability to identify their target, which may be expressions that appear prior in the text, abstractions of material that appeared earlier, or entities that exist independently from the text in existing

background knowledge. Seen in a more general way, the task is to cull out objects and events from multimedia sources (text, audio, video). An example challenge includes extracting entities within media and correlating those across media. For example this might include extracting names or locations from written/spoken sources and correlating those with associated images. Whereas commercial products exist to extract named entities from text with precision and recall in the ninetieth percentile, domain independent event extractors work at best in the fiftieth percentile and performance degrades further with noisy, corrupted, or idiosyncratic data.

Therefore work on the resolution of referring expression and the identification of entities in text and multimedia documents remains important fields of activity for the future.

Discourse and Dialogue

Extracting the knowledge contained in documents and understanding and generating natural dialog behavior requires more than the resolution of local semantic ambiguities. Intelligent analysis needs to consider the global argumentative structure of documents and discourse, and dialogs need to be analyzed for pragmatic content.

Computational work in discourse has focused on two different types of discourse: extended texts and dialogues, both spoken and written, yet there is a clear overlap between these two: dialogues contain text-like sequences spoken by a single individual and texts may contain dialogues. But application opportunities and needs are different. Work on text is of direct relevance to document analysis and retrieval applications, whereas work on dialogue is of import for human-computer interfaces regardless of the modality of interaction. Both are divisible into segments (discourse segments and phrases) with the meaning of the segments being more than the meaning of the individual parts.

The main focus of the research is the interpretation beyond sentence boundaries, the intentional and informational approach.

According to the informational approaches, the coherence of discourse follows from semantic relationships between the information conveyed by successive utterances. As a result, the major computational tools used here are inference and abduction on representations of the propositional content of utterances.

According to the intentional approaches the coherence of discourse derives from the intentions of speakers and writers and understanding depends on recognition of those intentions.

One difficulty is to build models of human-machine-dialog when initially only examples of human-human interaction exist, which may not be relevant. Bootstrapping suitable models will therefore require Wizard-of-Oz studies with simulated systems.

Natural Language Generation

In many of the applications mentioned above, systems need to produce high-quality natural language text from computer-internal representations of information. Natural language generation can be decomposed into the tasks of text planning, sentence planning and surface realization. Text planners select from a knowledge pool which information to include in the output and out of this create a text structure to ensure coherence. On a more local scale, sentence planners organize the content of each sentence, massaging and ordering its parts.

Surface realizers convert sentence-sized chunks of representation into grammatically correct sentences.

Generator processes can be classified into points on a range of sophistication and expressive power, starting with inflexible canned methods and ending with maximally flexible feature combination methods. It is safe to say that at the present time one can fairly easily build a single-purpose generator for any specific application, or with some difficulty adapt an existing sentence generator to the application, with acceptable results. However, one cannot yet build a general-purpose sentence generator or a non-toy text planner. Several significant problems remain without sufficiently general solutions:

- Lexical selection is one of the most difficult problems in generation. At its simplest this question involves selecting the most appropriate single word for a given unit of input. However as soon as the semantic model approaches a realistic size and as soon as the lexicon is large enough to permit alternative locutions, the problem becomes very complex. The decision depends on what has already been said, what is referentially available from context, what is most salient, what stylistic effect the speaker wishes to produce and so on. What is required: development of theories about and implementations of lexical selection algorithms, for reference to objects, events states, etc., and tested with large lexical.

- Discourse structure (see also there) So far, no text planner exists that can reliably plan texts of several paragraphs in general. What is required: Theories of the structural nature of discourse, of the development of theme and focus in discourse, and of coherence and cohesion; libraries of discourse relations, communicative goals and text plans: implemented representational paradigms for characterizing stereotypical texts such as reports and business letters; implemented text planners that are tested in realistic non-toy domains.

- Sentence planning: Even assuming the text planning problem is solved, a number of tasks remain before well-structured multi-sentence text can be generated: These tasks, required for planning the structure and content of each sentence, include: pronoun specification, theme signaling, focus signaling, content aggregation to remove unnecessary redundancies, the ordering of prepositional phrases, adjectives, etc. What is required: Theories of pronoun use, theme and focus selection and signaling, and content aggregation; implemented sentence planners with rules that perform these operations; testing in realistic domains.

- Domain modeling: a significant shortcoming in generation research is the lack of large, well-motivated application domain models, or even the absence of clear principles by which to build such models. A traditional problem with generators is that the inputs are frequently hand-crafted, or are built by some other system that uses representation elements from a fairly small hand-crafted domain model, making the generator's inputs already highly oriented toward the final language desired….What is required: Implemented large-size (over 10.000 concepts) domain models that are useful both for some non-linguistic application and for generation; criteria for evaluating the internal consistency of such models; theories on and practical experience in the linking of generators to such models: lexicon of commensurate size.

Probably the problem least addressed in generator systems today is the one that will take the longest to solve. This is the problem of guiding the generation process through its choices when multiple options exist to handle any given input.

The generator user has to specify not only the semantic content of the desired text, but also its pragmatic – interpersonal and situational – effects. Very little research has been performed on this question beyond a handful of small-scale pilot studies. What is required: Classifications of the types of reader characteristics and goals, the types of author goals, and the interpersonal and situational aspects that affect the form and content of language; theories of how these aspects affect the generation process; implemented rules and/or planning systems that guide generator systems' choices; criteria for evaluating appropriateness of general text in specified communicative situations.

Effective presentations require the appropriate selection of content, allocation to media, and fine grained coordination and realization in time and space. Discovery and presentation of knowledge may require mixed media (e.g., text, graphics, video, speech and non-speech audio) and mixed mode (e.g., linguistic, visual, auditory) displays tailored to the user and context. This might include tailoring content and form to the specific physical, perceptual, or cognitive characteristics of the user. It might lead to new visualization and browsing paradigms for massive multimedia and multilingual repositories that reduce cognitive load or task time, increase analytic depth and breadth, or simply increase user satisfaction. A grand challenge is the automated generation of coordinated speech, natural language, gesture, animation, non-speech audio, generation, possibly delivered via interactive, animated lifelike agents. Preliminary experiments suggest that, independent of task performance, agents may simply be more engaging/motivating to younger and/or less experienced users.

Ontologies

Large-scale ontologies are becoming an essential component of many applications including standard search (such as Yahoo and Lycos), e-commerce (such as Amazon and eBay), configuration (such as Dell and PC-Order), and government intelligence (such as DARPA's High Performance Knowledge Base program). As discussed in the preceding paragraphs, ontologies will constitute a major source of knowledge needed for several levels of NLP.

Ontologies are increasingly seen as an important vehicle for describing the semantic content of web-based information sources and they are becoming so large that it is not uncommon for distributed teams of people to be in charge of the ontology development, design, population, and maintenance.

Ontologies define a vocabulary for researchers who need to share common understanding of the structure of information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. The principal reasons to use an ontology in machine translation (MT) and other language technologies are to enable source language analyzers and target language generators to share knowledge, to store semantic constraints and to resolve semantic ambiguities by making inferences using the concept network of the ontology. An ontology contains only language independent information and many other semantic relations as well as taxonomic relations.

Though the utility of domain ontologies is now widely acknowledged in the IT (Information Technology) community, several barriers must be overcome before ontologies become practical and useful tools. One important achievement would be to reduce the time and cost of identifying and manually entering several thousand concept descriptions by developing

automatic ontology construction. Another important task is to find arrangements that make development and sharing of ontologies commercially attractive.

Some challenges for ontology research:

Work on ontologies needs to provide generally applicable top-ontologies that cover most important core concepts that will be needed for many domains. Extensions to new domains could then start by enriching these top-ontologies in a specific direction, reducing the initial effort for creating new ontologies, for merging independently developed extensions, and for rapid customisation of existing ontologies.

This requires that ontology-creators are willing to share parts of their work and find suitable processes to organize cooperation. It also requires the development of standards for the languages in which ontologies are specified and can be interchanged (e.g. along the lines of the OIL proposal). Here, the challenge is to find suitable compromises between expressive power and depth on one hand and ease of use on the other hand. Ideally, one specification language should be able to cover the whole spectrum up to advanced knowledge representation as used in the CYC project.

Incremental improvement of ontologies needs to be facilitated by specialized tools for easy visualization and modification. These tools (and the representations they work on) need to be domain-independent and suited even for casual users, and their design needs to be based on a user-centred process view.

It must be easy to plug in ontologies into various NLP-based tools such as tools for information extraction, organization and annotation of document collections (semantic Web), environments for terminology management and controlled language. This will permit to audit the contained knowledge in manifold ways, and will allow for rapid quality improvement.

What is required: tools that support broad ranges of users in (1) merging of ontological terms from varied sources, (2) diagnosis of coverage and correctness of ontologies, and (3) maintaining ontologies over time.

Lexicons

Lexical knowledge – knowledge about individual words in the language – is essential for all types of natural language processing. Developers of machine translation systems, which from the beginning have involved large vocabularies, have long recognized the lexicon as a critical (and perhaps the critical) system resource. As researchers and developers in other areas of natural language processing move from toy systems to systems which process real texts over broad subject domains, larger and richer lexicons will be needed and the task of lexicon design and development will become a more central aspect of any project.

A basic lexicon will typically include information about morphology and on the syntactic level, the complement structures of each word or word sense. A more complex lexicon may also include semantic information, such as a classification hierarchy and selectional patterns or case frames stated in terms of this hierarchy. For machine translation, the lexicon will also have to record correspondences between lexical items in the source and target language; for speech understanding and generation, it will have to include information about the pronunciation of individual words. For this purpose the overall lexicon architecture and the representation formalism used to encode the data are important issues.

No matter if we want to build an ontology or a lexicon, in general for this kind of high-quality semantic knowledge base, manual processing is indispensable. Traditionally computer lexicons have been built by hand specifically for the purpose of language analysis and generation. However, the needs for larger lexicons are now leading to efforts for the development of common lexical representations and co-operative lexicon development.

The area is ripe – at least for some levels of linguistic description – for reaching in the short term a consensus on common lexical specifications. We must expand the experiences with the sorts of semantic knowledge that could be effectively used by multiple systems. We must also recognize the importance of the rapidly growing stock of machine-readable text as a resource for lexical research. The major areas of potential results in the immediate future seem to lie in the combination of lexicon and corpus work. There's a growing interest from many groups in topics such as sense tagging or sense disambiguation on very large text corpora, where lexical tools and data provide a first input to the systems and are in turn enhanced with the information acquired and extracted from corpus analysis.

Machine Learning

As mentioned above, the acquisition of knowledge continues to impose on of the biggest difficulties to the application of NLP technologies. This holds both for linguistic knowledge (grammars lexicons) and for world knowledge (ontologies, facts). In order to make extensions of NLP to new domains possible, the acquisition process needs to be supported by algorithms that can exploit existing textual material and extract knowledge of various types from it.

Approaches to these methods can be found in various fields of research, such as statistical language models, bilingual alignment, grammar induction, statistical parsing, statistical classification technology, Bayesian networks and other ML methods used in artificial intelligence research, data mining techniques etc.

Due to the specific nature of lexical information, it is important to pick or develop methods that scale to large vocabularies and large sets of features and that can exploit multiple sources of evidence in a good way. Also, the methods need to be able to use a rich set of existing background knowledge, so that no effort is wasted in re-discovering what was already known.

It is important to have methods that can use richly annotated training data, but do not require that large datasets have to be annotated in this way. Instead, methods should be able to draw a maximum of advantage from raw data without annotation using unsupervised learning approaches. Also, it will be important to guide the effort of human annotation so that time is spent in the most efficient way, using active learning methods. Tools and processes for managing annotation projects (including assessment of quality levels) need to be developed and shared on a broad basis.

Whenever possible, one should try to use models that contain explicit linguistic representations (ideally organized along different strata) so that partial reuse of models and rapid adaptation to slightly different is facilitated.

4. Milestones

Some relevant items not included in Bernsen 2000.

Basic technologies

Short term

- accurate syntactic analysis for well-formed input from specific domains
- simple methods for minimizing annotation effort during domain adaptation
- ML algorithms that combine active and unsupervised learning for optimal exploitation of data
- generally applicable annotation schemes for semantic markup of text
- standards for encoding and exchange of ontological resources emerge
- top-level ontologies generally available
- tools for semi-automatic construction and population of ontologies from text
- tools for simple semantic enrichment of Web pages
- approaches to markup of discourse structure and pragmatics

Medium term

- improved methods for minimizing annotation effort during domain adaptation
- tools for adaptation of syntactic analysis to specific application with minimal human effort
- accurate syntactic analysis for slightly ill-formed input for restricted domains
- improved syntactic analysis of input with uncertainties (word lattices)
- machine learning methods that exploit and extend existing knowledge sources
- sufficiently accurate semantic analysis of free text from restricted domains
- generic schemes for the annotation of pragmatic content
- schemes for annotation of discourse and document structure
- generally usable ontologies exist for many domains
- NL generation verbalizes information extracted/deduced from multiple sources for QA
- Agent/user models for dialogs of moderate complexity

Long term

- accurate syntactic analysis for ill-formed input from multiple domains
- sufficiently accurate semantic analysis of free text from multiple domains
- recognition of pragmatic content in text and dialog
- NL generation produces stylistically adequate and well-structured text

Systems

Short term

- QA systems are able to answer simple factual questions
- Summarization system produce well-formed extracts from short documents
- automated e-mail response systems deliver high-quality replies in easy cases
- MT for information assimilation

Medium term

- QA systems that deduce answers from information in multiple sources
- Summarization systems are able to merge multiple documents
- Summarization systems are able to deliver different types of summaries
- Integration of translation memories with MT enables fast domain-adaptation
- Mixed-initiative dialogue systems for services and e-commerce

Long term

- Translator's workbenches based on TM, MT, and multi-modal input facilities
- QA systems that are able to explain their reasoning

5. Recommendations for NLP research in Europe

1. Build and make publicly available at low cost large-scale multilingual lexical resources, with broad coverage, generic enough to be reusable in different application frameworks
2. To turn special attention to the development of better ontologies which are reusable across domains in order to encode static world knowledge
3. Creation of large common accessible multilingual corpora of syntactical and semantically annotated data annotated also beyond sentence boundaries

4. Encourage development of statistical and machine-learning methods that facilitate bootstrapping of linguistic resources

5. Common standards will improve the effectiveness of people's cooperation, the identification of the requirements for the system specification, the inter-operability among systems and the possibility of re-using and sharing system components.

6. Integration of language processing into the rest of cognitive science, artificial intelligence and computer science e.g. some ambitious projects centered on NL but combining various techniques and different areas of AI. New type of projects: Very different for scale, ambition and timeframe

7. Establishment of centers of excellence as focus points for projects for a period of five to ten years.

8. Encourage systematic evaluations (but how ?)

6. References

- Berners-Lee, T. (2001) The Semantic Web, Scientific American (5/2001)

- Bernsen , N.O. (2000) Speech-Related Technologies. Where will the field go in 10 years? roadmap workshop, Katwijk

- Burger, J. e.a. (2000) Issues, Tasks and Program Structures to Roadmap Research in Question & Answering, Memo National Institute of Standards and Technology, Gaithersburg

- Carbonell, J. e.a. (2000) Vision Statement to Guide Research in Q&A and Text Summarization, Memo National Institute of Standards and Technology, Gaithersburg

- Cole,R.A. (Ed.). (1997) Survey of the State of the Art in Human Language Technology Cambridge University Press, Cambridge

- Declerck, Th., Wittenburg, P., Cunningham, H. (2001) The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment, ACL Workshop, Toulouse

- Delannoy, J.-F. (2001) What are the points? What are the stances? Decanting for question-driven retrieval and executive summarization, ACL Meeting, Toulouse

- Fensel, D. Hendler, J., Lieberman, H. ,Wahlster, W. (2000) Dagstuhl-Seminar: Semantics for the WWW, Dagstuhl, Germany

- Grishman, R. and Calzolari, N. Lexicons in Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge

- Grosz, B. (1997) Discourse and Dialogue in Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge

- Heisterkamp, P., (2000) Speech Technology in the year 2010, roadmap workshop, Katwijk
- Hirschman, L. and Thompson, H.S. (1997) Evaluation in Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge
- Hovy, E., (1997) Language Generation in Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge
- Kang, S.-J. and Lee, J.-H. (2001) Semi-Automatic Practical Ontology Construction by using Thesaurus, Computational Dictionaries, and Large Corpora, ACL workshop Toulouse
- Kay, M. (1997) Machine Translation: The Disappointing Past and Present. In: Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge
- Kay, M. (1997) Multilinguality. In: Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge
- Knight, K. (2001) Language Modeling for Good Generation, Workshop on Language Modeling and Information Retrieval, Pittsburgh
- Krauwer, St., (2000) Going from 'what' to 'why' across language barriers in the unified distributed information space. Roadmap workshop, Katwijk
- Maybury, M.T. and Mani, I., (2001) Automatic Summarization, ACL Meeting Toulouse
- Maybury, M.T., (2001) Human Language Technologies for Knowledge Management: Challenges and Opportunities, ACL Meeting, Toulouse
- Pardo, J.M., (2000) How will language and speech technology be used in the information world of 2010? Research challenges & Infrastructure needs for the next ten years. Report on the Roadmap Workshop, Katwijk aan Zee
- Staab, St., (2001) Knowledge Portals, ACL Meeting, Toulouse
- Stock, O. (2000) Processing Natural Language from 2000 to 2010, roadmap workshop, Katwijk
- Velardi, P. and Missikoff, M. and Basili, R. (2001) Identification of relevant terms to support the construction of Domain Ontologies, ACL workshop Toulouse
- Uszkoreit; H. (2001) Crosslingual Language Technologies for Knowledge Creation and Knowledge Sharing, Toulouse
- Zaenen, A. and Uszkoreit, H. (1997) Language Analysis and Understanding. In: Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge