Methodologies for Knowledge Acquisition from NL Texts

Costanza Navarretta
Center for Sprogteknologi
Njalsgade 80
2300 Copenhagen S
email: costanza@cst.ku.dk

1 Introduction

The MECKA project, "Methodologies for Constructing Knowledge Bases for NLP Systems", is funded by the EU and involves HCRC - LTG Edinburgh with Andrei Mikheev and Marc Moens (supervisor), CST Copenhagen with Annelise Bech (supervisor and coordinator) and Costanza Navarretta, SRI-International California with Jerry Hobbs (expert consultant). The contents of this paper reflect the results of the joined work of all these people, but the author is of course responsible for any possible errors.

In the field of knowledge engineering for natural language the interest has been more centered on the representation language for knowledge bases than on methodological issues; no general method for constructing knowledge bases for NLP systems has yet been defined. The main objectives of the MECKA project are to define a sound and general methodology for constructing knowledge bases for natural language understanding systems and to look at the reusability potential of existing linguistic resources. The point of departure for the present investigation is which knowledge needs to be extracted and how it can be acquired. The ultimate goal is to make the construction of knowledge bases for NL systems less labour-intensive, to improve the quality and the reliability of the resulting knowledge bases and to examine the possibility of automatising (parts of) the methodology.

The knowledge we are interested in is the background knowledge which is presupposed by the texts. Background knowledge can be divided into *commonsense knowledge* (or *domain independent knowledge*) and *domain specific knowledge.* The granularity of the knowledge to be extracted depends on the domain of discourse and on the task of the system. Background knowledge comprises *linguistic* (*word*) and *extra-linguistic* (*world*) knowledge. Because the two kinds of knowledge are strictly interrelated, we will not try to establish a clearcut line between them but, when possible, we will indicate which kinds of linguistic knowledge require extra-linguistic knowledge to be disambiguated.

The knowledge acquisition process comprises knowledge elicitation and knowledge organization. In this paper we concentrate on the former.

2 Different strategies for constructing knowledge bases

In the first phase of the project we made a state-of-the-art survey of the different strategies which have been applied for constructing knowledge bases for NLP systems. We identified the following tendencies: top-down vs. bottom-up strategies, knowledge-based vs. lexicalist systems, corpus-supported vs. expert-based approaches.

2.1 Top-down vs. bottom-up strategies

Top-down strategies start off from pre-defined, non-linguistic characterization of knowledge structures, the top layers of the ontology, and then relate the knowledge needed by the actual texts to these top layers.

The pre-defined knowledge structures provide a means for ordering the knowledge elicited from the texts so that the knowledge base can be constructed in a consistent way. A drawback of top-down strategies is that the knowledge structures used are often not completely adequate to represent all the knowledge required by the texts.

Bottom-up strategies start from linguistic expressions which must then be ordered in a constantly evolving model. The model is hence always adequate to the knowledge to be represented, but it is difficult to organize the elicited knowledge in a consistent way, especially when dealing with large text corpora.

The two methods should be combined so that the bottom-up growth is alternated with some kind of top-down design.

2.2 Knowledge-based systems vs. lexicalist systems

Knowledge-based systems often have a relatively underspecified lexicon and a rich knowledge base with general and domain specific rules and an inference engine. More recently a lot of research has been done in order to create general lexicons for natural language processing systems which contain a lot of commonsense knowledge, so that this knowledge can be (re)used by different systems/applications.

We think that the distinction between knowledge-based systems and lexicalist systems is more historical than methodological, the work on the creation of knowledge-rich and large lexicons being relatively new.
The studies made in the field of lexical semantics are very relevant to our work because they contribute to the specification of recurrent phenomena that presuppose world knowledge and we are investigating to what extent methodologies for constructing knowledge bases can be facilitated by incorporating some of the new lexicons.

2.3 Corpus-based and expert-based approaches

In corpus-based approaches the primary knowledge source is a large general and/or technical text corpus. Very few researchers have used this approach because it is extremely resource-consuming. However the interest for large text corpora is growing in all fields of computational linguistics. We contrast the corpus-based approach with the expert-based approach, where information of different nature (i.a. acquired from experts in the actual domain or from common people as "experts of the language") is used. There is an intermediate approach, which involves the uses of other linguistic resources as dictionaries and term banks, which are constructed by lexicographers and terminologists using knowledge extracted from text corpora.

We think that the reliance on human experts when acquiring knowledge from texts can be reduced by using corpus-supported and intermediate approaches.

3 A corpus-based knowledge methodology

We have decided to work on Hobbs' three-step strategy which was developed for the TACITUS system (Hobbs 1984) because it is corpus-based, it combines bottom-up growth with an amount of top-down design and can be extended to include supporting linguistic resources.

Hobbs' three-step strategy is the following:

1. Select the facts that should be in the knowledge base, by determining which facts are linguistically presupposed by the texts.
2. Organize the facts into clusters and within each cluster, according to the logical dependencies among the concepts they involve.
3. Encode the facts as predicate calculus axioms, regularizing the concepts, or predicates, as necessary.

Though it is promising the above strategy is too underspecified and it relies too much on the intuition of the single knowledge engineer. Applying the first two steps of the strategy (dealing with knowledge elicitation) to small text corpora from different domains we have come to the following refined method[1]:

1. List the content words in the text corpus to be processed and make a list of general relevant facts about the text corpus and about the content words in it.

2. Group morphologically related words.

3. Divide the resulting groups of morphologically related words into subdomains.

4. Give a first organisation of the knowledge in each subdomain.

[1] Of course, we presuppose, as Hobbs does, that the knowledge engineer has acquired general knowledge about the actual domain before applying the elicitation method on a text corpus.

Next for each content word (or for each group of morphologically related words) do:

a. Look for all occurrences of the word in the text corpus to see the contexts in which the word is used. When necessary, look at previous or following sentences to resolve anaphora.

b. Reduce the occurrences to their predicate-argument relations. Examine the contexts and determine what facts about the word are required to justify each of its occurrences.

c. Make a preliminary division of these predicate-argument relations into heaps, according to a first analysis of which predicates should go together. Patterns should be split up when more facts are presupposed in the actual citation.

d. Give an abstract characterization of the facts about the word that justify each of the heaps. Recognizing a more abstract characterization may lead to joining of two heaps and failure to find a single abstract characterization may lead to splitting a heap.

3.1 An example

In the following a simplified example of the application of the methodology to a short extract from our "car owner's manual" corpus is given.

Never tow an automatic transmission model with the rear wheels raised (with the front wheels on the ground) as this may cause serious and expensive damage to the transmission. If it is necessary to tow the vehicle with the rear wheels raised, always use a towing dolly under the front wheels.

A first coarse-grained division of the content words into domain-independent and domain-specific knowledge is the following:

Domain independent knowledge:

always, cause, damage, expensive, front, ground, may, necessary, never, raise, rear, serious, under, use.

Domain specific knowledge:

dolly, model, tow, automatic transmission, vehicle, wheel.

Then the words are divided into domains and subdomains. In this example the "clusters of commonsense knowledge" described in Hobbs et al. (1986) have been used:

under
front
rear===>orientation

ground===> space, orientation

tow
raise ===> space, movement, causality

necessary
cause ===> causality

always
never ===> time

expensive ===> scale, economics

damage ===> causality, goal-directed systems,
functionality

serious ===> scales

aut. transmission
vehicle
wheels
dolly
model ===> artifacts, goal-directed-systems

For analysing *damage* we extracted the occurrences of damage and of morphologically related words from the entire text corpus. Some of the extracted occurrences are the following:

Caution: Anti-freeze will damage paintwork.

Incorrect towing equipment could damage your vehicle.

Never tow an automatic transmission model with the rear wheels raised (with the front wheels on the ground) as this may cause serious and expensive damage to the transmission.

Note that alloy wheels use special nuts incorporating a washer to prevent damage to a roadwheel.

The wheels and axle on the ground must be in good condition. If they are damaged, use a towing dolly.

> Operating with insufficient amount of oil can damage the engine,
> and such damage is not covered by warranty.

In the subsequent step of reducing the extracted occurrences to their predicate-argument relations, we found out that we could use the pattern”for X to cause damage in Y is for X to damage Y".

Then the resulting predicate-argument relations were divided into the three heaps:

1)
overfilling damages engine
antifreeze damages paintwork
insufficient oil damages engine
driving with deflated tire damages tire
incorrect towing damages transmission...

2)
prevent damage
ensure no damage
risk damage...

3)
serious damage
expensive damage
damage beyond repair
damage covered by warranty...

The resulting characterization (generalization) for each of the three heaps is:

1. incorrect procedure damages component
2. damage is bad
3. damaged components need to be repaired, and repairs cost money.

4 Further improvements of the methodology

The defined methodology for knowledge elicitation can be applied to texts from different domains and can be partially automated (using e.g. KWIC tools and tools for clustering morphological related words). However it still requires extremely many resources and it relies too much on the intuition of the knowledge engineer who applies it. There are different possible improvements, e.g. making implicit knowledge explicit and shoring up the knowledge elicitation process with existing linguistic resources such as large general language and technical corpora.

4.1 Making implicit knowledge explicit

To improve the process of making explicit the knowledge which is implicit in the texts, it is necessary to consider both text-level and word-level information and to systematize the linguistic carriers of background information.

The first step is to determine the system task (which is i.a. necessary to define the appropriate granularity of knowledge) and to analyse the actual text corpus. The latter point comprises looking at a) the general knowledge about the corpus which clarifies many linguistic and pragmatic features of the texts (e.g. linguistic conventions of the genre, style, medium, discourse strategy, informational density, length and complexity of the sentences, types of subordinate clauses, overt and covert connections among sentences, temporal and causal relations about sentences); b) knowledge about the communicative situation and communicative competence (the addressor and the addresse of the texts and their qualifications, the purpose of the communication, the extent of shared knowledge etc.) which is relevant to determine the granularity of the domain (degree of technicality) and to establish many facts presupposed by the texts (i.a. the purpose of the texts).

All domains (but with different granularity) require knowledge about scales, physical objects, space, change, causality, time, functionality etc.[2]

The dispute whether linguistic and extra-linguistic knowledge should be handled in distinct ways is still not resolved. We will not try to define a clearcut division line between the two kinds of knowledge which in most cases are strictly interrelated, but, when possible, we will determine the linguistic phenomena that indicate the presence of presupposed background knowledge. This makes it possible to investigate regularities and dependencies between the two kinds of information because most of the facts presupposed in the texts are exactly the facts that are necessary to resolve/disambiguate linguistic phenomena. Among the phenomena which in many cases 'indicate' implicit background knowledge are compound nouns, pronominal and nominal anaphora, definite reference, attachment ambiguity, metonymy, ellipses, metaphors, belief reports. In the example '*Never tow an automatic transmission model with the rear wheels raised (with the front wheels on the ground) as this may cause serious and expensive damage to the transmission.'* it is necessary to have access to knowledge about cars and kinds of cars to resolve the definite reference of 'the rear wheels' (here a bridging reference). For disambiguating the compound nominal 'automatic transmission model' one must know that in the actual context a model is a particular make of car and that there is a connection between front wheels and automatic transmission.

4.2 Typologically specialising the methodology and using large text corpora

To make the knowledge elicitation process more automatable and then less labour-intensive, we have investigated strategies and techniques to facilitate

[2] These are the "clusters of commonsense knowledge" identified by most researchers in the fields of knowledge engineering (Hayes 1979, Herzog and Rollinger (eds.) 1991, Hobbs and Moore (eds.) 1985, Lenat and Feigenbaum 1987).

the extraction of specific types of knowledge. For this purpose we have determined to which extent it is possible to incorporate in our methodology statistical strategies which have been applied for i.a. identifying terms in technical corpora (Huizhong 1986), clustering words according to their distribution in large text corpora (Pereira et al. 1993, Hatzivassiloglou and McKeown 1993) and disambiguating word senses (Justeson and Katz 1993).

We believe that the quality of the information extracted for constructing knowledge bases in many cases can be improved by using large general and technical text corpora as support material. We are also interested in the above strategies because they give the possibility of automatically sorting the huge amount of information contained in large text corpora. Moreover, some of the statistical strategies considered address the problem of sparseness of data, showing a promising way for supporting the knowledge elicitation process and for improving the reliability of the resulting knowledge bases.

One problem with many of the statistical strategies we have looked at is that they are quite new and have only been applied experimentally, thus their results must be taken with reservation until they are validated.

4.3 Reusing existing resources

To reduce the costs related to the process of acquiring knowledge from texts and to improve the reliability of the resulting knowledge bases we are also investigating the reusability potential of existing linguistic resources. The resources which we have found most promising are machine-readable dictionaries, term banks, lexical bases and tools for processing large text corpora. At present we are looking at these resources and are defining how they can support our methodology.

5 Conclusions

We have discussed some general tendencies in the field of knowledge engineering for natural language processing systems. We have defined a general elicitation methodology for natural language understanding systems refining Jerry Hobbs' three-step strategy. The methodology is linguistically anchored and can be applied to very different domains, but it still presents some problems, i.e. it is very time-consuming and it relies too much on the skills of the single knowledge engineer. At present we are working to obviate these problems to improve the viability of the knowledge elicitation process and the reliability of the elicited material.

6 References

Goodman, K. and S. Nirenburg (eds.), 1991. *The KBMT Project: A Case Study in Knowledge-Based Machine Translation*. San Mateo, Ca., Morgan Kaufmann.

Hatzivassiloglou, V. and K. R. McKeown, 1993. 'Towards the Automatic

Identification of Adjectival Scales: Clustering adjectives according to meaning.' In: *ACL Proceedings, 31st Conference*, Columbus, Ohio, USA, 1993, pp. 172-182.

Herzog, O. and C.-R. Rollinger (eds.), 1991. *Text Understanding in LILOG.* Berlin. Springer-Verlag.

Hobbs, J.R., 1984. *Sublanguage and Knowledge.* Technical Note 329. SRI, California.

Hobbs, J.R. and R.C. Moore (eds.), 1985. *Formal Theories of the Commonsense World*. Ablex, New Jersey.

Hobbs, J.R., W. Croft, T. Davies, D. Edwards, K. Laws, 1986. *Commonsense Metaphysics and Lexical Semantics*. Technical Note 392. SRI, California.

Hobbs, J.R., M. Stickel, D. Appelt and P. Martin, 1990. *Interpretation as Abduction*. Technical Note 499. SRI, California.

Huizhong, Y., 1986. 'A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts. (An Interim Report)', in *Literary and Linguistic Computing*, Vol. 1, no. 2, pp. 93-103.

Justeson, J.S. and S.M. Katz, 1993. 'Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns'. In: *Making Sense of Words, Proceedings of the 9th Annual Conference of the UW Centre for the new OED and Text Research*. September 1993, Oxford England, pp. 57-73.

Lenat, D.B. and E.A. Feigenbaum, 1987. *On the thresholds of Knowledge.* MCC Technical Report Number AI-126-87, Austin Texas.

Lenat, D.B. and R.V. Guha, 1988. *The World According to CYC*. MCC Technical Report Number ACA-AI-300-88, Austin, Texas.

Pereira, F.P., N. Tishby and L. Lee, 1993. 'Distributional Clustering of English Words'. In: *ACL Proceedings, 31st Conference*, Columbus, Ohio, USA, pp.183-190.

Pustejovsky, J., 1991. 'Towards a Generative Lexicon'. In: *Computational Linguistics*, Vol. 17, no. 4, pp. 409-441.