Annotating a corpus for building a domain-specific knowledge base.

Sabine Bartsch
Darmstadt University of Technology, Department of Linguistics and Literature
Hochschulstr. 1, D-64289 Darmstadt, Germany
e-mail: bartsch@linglit.tu-darmstadt.de

Abstract

The project described in this paper seeks to develop a knowledge base for the domain of data processing in construction - a sub-domain of mechanical engineering - based on a corpus of authentic natural language text. Central in this undertaking is the annotation of the relevant linguistic and conceptual units and structures which are to form the basis of the knowledge base. This paper describes the levels of annotation and the ontology on which the knowledge base is going to be modelled and sketches some of the linguistic relations which are used in building the knowledge base.

1 Introduction

Texts are a prime medium for the dissemination of information in the sciences. In this respect, bodies of text of a specific domain serve the function of an information store collectively representing the knowledge of a target domain. The human language user can tap the information contained in texts representing the knowledge of a domain by means of an intractable, complex intellectual process, yet other paths to the exploitation of the information represented in texts are required in order to manage and process the ever growing amount of text in many specialized domains (especially in domains where information is growing rapidly such as e.g. bio-technology). To this end, strategies for accessing the information stored in texts have to be developed that complement and model the intellectual capacity of the human language user. Alternative modes of access to the information stored in texts are required in order to allow human language users as well as computational systems the exploitation of the stored information. In order to achieve this aim, different representations, e.g. indices, ontologies, knowledge bases etc., have to be devised in order to represent information for different requirements and agents. As texts are encoded in natural language, natural language processing based on large corpora plays an important role in this task (Bateman, 1992), but in order to exploit the information contained in text corpora for the creation of knowledge bases, the organization and distribution of information in texts has to be analysed and made explicit by means of appropriate textual annotation.

The basic aim of the project sketched in this paper is the development of techniques for the exploitation of natural language corpora as knowledge bases by transferring the information provided by linguistic and conceptual text structures to other forms of representation, thus allowing both the human user and computational systems alternative and systematic ways of access to the information contained in texts. The target domain of this project is the domain of data processing in construction, a sub-discipline of mechanical engineering. This domain is interesting in knowledge representation terms itself because it is concerned with representations of information in different modes of representation (natural language, object oriented programming, visual representations e.g. in CAD modelling etc.).

In the project, annotations and tools suitable for the exploitation of a bi-lingual (English, German), multi-media (natural language, OO programming, visual representations, CAD models etc.) corpus are being developed. This paper reports on the annotation of the information deemed interesting for the planned corpus application. The structures and clues to be exploited range from the individual lexical item and its characteristic co-occurrences and collocations to the linguistic structures and relations establishing the overall textual context.

2 The nature of text as a knowledge base

It is an important observation in the study of language that meaning and linguistic structure are closely connected with one another (Hunston & Francis, 1999: 83; Sinclair, 1991: 65). This observation is of central importance when it comes to answering questions about the correlations between the information contained in text and its representation by means of linguistic structures.

In order to exploit a corpus of natural language text as a knowledge base, information has to be identified in text based on linguistic and conceptual structures. Additional problems are introduced by the fact that apart from natural language text, the corpus under study also includes resources in other media, e.g. Java programming code, visual representations in the form of pictures, diagrams and CAD models. The following paragraph will first and foremost discuss some central aspects of the nature of natural language texts and then add some remarks about the incorporation of the other types of "text" represented in the corpus under study.

Information is stored in natural language texts (a) in form of the lexical items representing the concepts, (b) in form of the conceptual relations relating the concepts of a specific domain to each other, and (c) by means of the linguistic structures within which they are embedded. The first aspect is generally addressed in projects focussing on terminology, the second aspect is the focus of conceptually-based ontology projects (e.g. On-To Knowledge[1]). The third aspect, viz. the role of linguistic structure, is a much neglected aspect in knowledge representation; it is the focus of the present project. All three aspects interact and must, consequently, be observed simultaneously in order to develop appropriate models for the representation of information in natural language texts.

[1] On-To Knowledge http://www.ontoknowledge.org/about.html

The interaction of these structural levels contributes to the network character of information in texts, yet at the same time raises some of the central issues in the context of attempts at exploiting natural language texts as knowledge bases. These issues can be identified as follows:

(1) identification of units of information;
(2) identification of the distribution of these units relative to one another;
(3) identification of the organization of information;
(4) identification of linguistic structures;
(5) identification of the relevant coincidences of (1)–(4) in order to point out the correlation between linguistic structure and organization of information in text.

Issue (5) is in keeping with the underlying tenet of this project that if language is the prime medium for the dissemination of information then there must be a correlation between the organization of information and linguistic structure. Investigating this correlation and determining the functional contribution of linguistic structure to the organization of information is a central project aim.

As far as the other media represented in the corpus are concerned, these will be analysed (a) regarding their unique contribution to the organization of information across the texts within which they are embedded, and (b) regarding their unique organization of information and contributions to the representation of information in the target domain. This part of the corpus analysis will be described in a future paper.

At this point, some remarks about the status of the textual resource at the centre of the project may be called for. While many existing projects can build on previously existing domain specific databases (e.g. use of the MeSH thesaurus[2] in the MuchMore project[3]), such resources do not exist for the domain this project seeks to model. Also, the approach of taking unannotated corpora as a starting point and building the system from scratch has many advantages if it is to be applied to a broader range of domains: (a) There are few domains in which there already exist sufficient amounts of pre-coded information resources; (b) by building knowledge bases on already existing databases, ontologies and taxonomies, a certain bias may be introduced that could be undesirable as structures and phenomena occurring in real-life corpora might be overlooked. The "knowledge base built on corpus"-approach taken in this project has the advantage of building systematically on a well-defined source of information, i.e. the corpus, and can be extended in a systematic fashion at any time by expanding the corpus.

3 Corpus encoding: Levels of annotation

The corpus is completely encoded in *XML*, an encoding choice which is evolving into standard practice in corpus and computational linguistics (cf. the TEI guidelines for corpus encoding, XCES, etc.)[4]. It is also paramount that tools be used for linguistic analysis and annotation (pos tagging etc.) that either produce XML compliant output or whose output can be mapped onto the XML annotation structure of the corpus. Most of the tools currently available for linguistic analysis and annotation necessitate the implementation of the latter solution because they do not produce XML-compliant output. Also, there are, as yet, few ready-made tools or dedicated programming languages available for XML corpus inspection and further analysis. These issues will have to be addressed in the course of the project in order to find scalable solutions that will allow for the desired performance and usability when it comes to the implementation of larger, dynamic corpora.

3.1 Corpus pre-processing: Tokenization

In a first step, the corpus texts are *tokenized* at the text, sentence and word level. The corpus is broken down into paragraphs, sentences and individual lexical items. This step is an important prerequisite for the concurrent analysis and annotation steps.

The tokenizer currently used is Qtoken, a portable, Java tokenizer written by Oliver Mason that can split the corpus into individual lexical items, punctuation marks etc. The remainder of the tokenization process, i.e. the mark-up of paragraphs and sentences, is done by simple script processing.

3.2 Syntactic annotation

A variety of linguistic processing tools are run over the corpus in order to establish basic linguistic structures and attain a sufficient level of linguistic annotation that will allow the analysis of more complex linguistic structures.

The corpus is *part-of-speech tagged* with Qtag[5], a portable probabilistic part-of-speech tagger also by Oliver Mason. The output tagset is a variant of the Brown/Penn-style tagsets and has been agreed upon by Lancaster and Birmingham for a joint project on tagger evaluation. The output is mapped into attributes to the individual lexemes in the XML files.

The corpus is also *chunked* and *parsed* for shallow syntactic structure in order to enable the identification of dependency and argument structure. An example of the mark-up is shown in Figure 1 below.

3.3 Lexical and collocational relations

The corpus is analysed from the lexical level upwards starting with the features of and relations between individual lexical items in a text, extending to syntagmatic relations such as characteristic collocations and to the level of argument structure etc.

In a first step, the corpus is analysed for cohesive relations between lexical items. *Lexical cohesion* is an important device for making texts 'hang together' and provides important clues to the general topic flow of a text (cf. Halliday & Hasan 1976). A simple cohesive device is the repetition of a lexical item across a text, either by simple repetition of a particular lexical item or by repetition of other inflectional or derivational forms of a lexical item. Other, more complex forms of lexical cohesion build on complex semantic relations between the words in a text such as e.g. synonymy, hypernymy, hyponymy, antonymy and meronymy (Halliday & Hasan, 1976). Analysis of

[2] The National Library of Medicine's Medical Subject Headings (MeSH): http://www.nlm.nih.gov/mesh/meshhome.html

[3] MuchMore http://muchmore.dfki.de

[4] TEI Guidelines for Corpus Encoding http://www.tei-c.org/; XCES Corpus Encoding Standard for XML http://www.cs.vassar.edu/CES/, part of the EAGLES Guidelines.

[5] Qtag: http://web.bham.ac.uk/o.mason/software/tagger/; Qtoken: http://web.bham.ac.uk/o.mason/software/tokeniser/

lexical relations provides important information e.g. on synonymous or otherwise related lexical items and their syntactic-semantic features. By this approach, the verb classes employed in the identification of logical structures for the knowledge base (cf. 4.2 below) can be extended as required for a particular corpus.

Part of the corpus annotation will mark up the devices and structures creating lexical cohesion. This part of the mark-up will be based on the lexical relations encoded in WordNet[6] (Fellbaum, 1998), a free lexical database. Fankhauser & Teich (2004) describe an XML/XSLT solution based on WordNet for marking up lexical cohesion in text which relies on a semantically tagged corpus (SemCor). Pending tests will have to show whether a similar solution can be made to work with semantically unannotated corpora or sparse, non-domain specific semantic mark-up.

Collocations, i.e. characteristic co-occurrences of lexical items, give important clues to their contextual meaning as well as their contribution to the general topic flow of the text. True to Firth's (1957) dictum that "you shall know a word by the company it keeps", characteristic co-occurrences of lexical items with other lexical items are an important feature of a coherent text and contribute largely to its natural and native-like character. Another important feature of collocations that is especially relevant for domain-specific text is that many lexical items tend to contract specific collocates when they occur in a particular domain specific meaning (e.g. "torque is transmitted", "embodiment design") such that the characteristic collocations a lexical item contracts give important clues to the domain specific nature of a text and the information contained therein.

In this project, collocations are identified statistically mainly based on the Mutual Information score which is an information theoretic measure of association between co-occurring lexical items. Hindle (1994: 123, footnote) defines the MI score as a means "to identify relations that occur more often than chance, comparing the probability of observing word x and y together (the joint probability) with the probability of observing x and y independently."

The following formula from Church et al. (1991: 120 ff.) is applied in this project:

$$MI(x; y) = \log_2 \frac{p(x, y) \times N}{p(x) \times p(y)}$$

where p is the probability of the occurrence of the lexical items x and y and N is the sample size. The MI score provides a measure of the amount of information that occurrence of one item yields about the expected co-occurrence with the other as opposed to independent co-occurrence. In the annotated corpus, each running lexical item is given a unique identifier introduced as an attribute id="running number" by which all collocations contracted by a lexical item can be identified and made accessible in the corpus representation.

It is well-known that particular *grammatical patterns* are associated with particular meanings (cf. e.g. Hunston & Francis, 1999 pattern grammar approach; Levin, 1993 verb class alternations). At the level of syntax, the analysis focuses on specific grammatical patterns that are characteristically used to convey particular types of information. As an example, structures of advice and obligation (marked e.g. by the modal *must*) are employed heavily in instructional parts of the corpus:

Ex. 1: During the embodiment phase, […], designers must determine the overall layout design.

Ex. 2: […], the definitive layout must be developed to the point […].

```
<?xml version="1.0" encoding="UTF-16"?>
<document title="Embodiment design">
<paragraph>
<sentence no=″00052″
<wf pos="CS" id=″0011″> For</wf>
<wf pos="NN" id=″0012″> instance</wf>
<wf pos="," id=″0013″> ,</wf>
<wf pos="VBG" id=″0014″ vbcl=″11.1″>transmitting</wf>
<wf pos="NN" id=″0015″> torque</wf>
<wf pos="CC" id=″0016″> and</wf>
<wf pos="VBG" id=″0017″ vbcl=″29.5″> allowing</wf>
<wf pos="CS" id=″0018″> for</wf>
<wf pos="JJ" id=″0019″> radial</wf>
<wf pos="NN" id=″0020″> movement</wf>
<wf pos="IN" id=″0021″> by</wf>
<wf pos="NN" id=″0022″> means</wf>
<wf pos="CS22" id=″0023″> of</wf>
<wf pos="DT" id=″0024″> a</wf>
<wf pos="JJ" id=″0025″> flexible</wf>
<wf pos="NN" id=″0026″> shaft</wf>
. . .
</sentence>
</paragraph>
</document>

<collocations>
  <colloc id=″colloc003″ const1=″0014″ const2=″0015″>
</collocations>
```

Figure 1: Sample of the mark-up

4 Building the Knowledge Base

In order to attain the goal of building a knowledge base based on the correlations between linguistic structure and organization of information in the corpus texts, the lexical items and structures identified in the linguistic annotation have to be viewed and classified against the backdrop of a linguistically based ontology. For this purpose, the project draws on the classification set up by the Penman Upper Model (Bateman et al., 1989; Bateman, 1990)[7].

4.1 The Penman Upper Model

The vantage point of this investigation is linguistic structure, therefore the project builds on the Penman Upper Model for ontology development because it is one of few linguistically based ontologies (cf. also Corelex[8]). It suits the purposes of the project because it was developed to mediate between domain knowledge and natural language systems (Bateman, 1990). The aim of the Penman Upper Model is to introduce a concept according to its impact on the choice of grammatical constructions, lexical expressions etc. that can express it (cf. Bateman et al., 1989; Halliday & Matthiessen, 1999). As the project is interested in the differences between knowledge representation in different modes of representation and the vantage point from which the project is looking at knowledge representation is a linguistic one, viz. looking at how natural language represents knowledge, it is self-explanatory that linguistic structure is the main focus of the project.

[6] WordNet: http://www.cogsci.princeton.edu/~wn/

[7] There is also a multi-lingual extension to the Penman Upper Model, the Generalized Upper Model (cf. http://www.uni-bremen.de/~bateman/

[8] http://www.cs.brandeis.edu/%7Epaulb/CoreLex/corelex.html

The Penman Upper model recognizes e.g. such structures representing material processes of directed action,

Ex. 3: “torque is transmitted by a flexible shaft”,

which are frequent and central in texts of the target domain. Linguistic structures found in the corpus will be classified according to such categories as proposed in the Penman Upper Model.

4.2 Logical structures

In order to determine the relations between participants and objects across the overall textual information structure, logical structures and relations are annotated. Based on the syntactic-semantic classes of verbs (for the classification cf. Levin, 1993) instantiated in the corpus, more detailed analyses of logical structures of sets of verbs occurring in similar grammatical patterns are carried out. The verbs and their characteristic patterns provide the central basis for establishing the types of relations between the candidate items for the knowledge base. The resulting relations are of the type:

- “is transmitted by”
  ⇨ “torque is transmitted by a shaft”
- “material”
  ⇨ “is made of wood”
- “is a result of”
  ⇨ “torque results from the motion of a shaft”

This level of analysis builds on the combination of verb semantics and characteristic grammatical patterns into which these verbs enter and in which they attract specific types of participants. Constraints on domain-specific collocation patterns in the argument structure are identified by means of the aforementioned collocation analysis (cf. 3.3 above).

By analysing the correlation between the syntactic-semantic properties of verbs occurring in particular grammatical patterns and their place in the Penman Upper Model classification, the project seeks to build the knowledge base. The knowledge base is based on the correlation between the lexical items occurring in the text, their patterns of co-occurrence - (a) within specific syntactic patterns, and (b) with a specific set of other lexical items, - and their role within the general organization of the text structure. Thus, a sentence such as Ex. 3 “torque is transmitted by a flexible shaft” above, would be analysed along the following lines:
(1) ‘transmit’ is classified as a verb of ‘transmission and transfer’ (corresponding to Levin verb class 11.1),
(2) ‘transmit’ is the main verb in a passive construction;
(3) ‘transmit’ contracts a domain-specific collocation with the noun ‘torque’;
(4) the structure is classified in the Penman Upper Model as a material process of directed-action; and has the
(5) logical structure: x transmits y, where x is an inanimate agent and y is a directed force or motion.
The logical relations between the participants established by the verbs and the grammatical structures within which they are embedded form the basis for the relations between items in the knowledge-base.

5 Discussion

The annotation steps described in this paper pursue the aim of preparing a corpus of naturally occurring text as a resource for building a knowledge base. In order to achieve this aim, a number of issues have to be addressed in the context of discovering the distribution and organization of information in natural language text and the appropriate levels of analysis and annotation for different modes of representation (indices, ontology, taxonomy, structural display, information retrieval etc.). Questions yet to be answered are whether the planned levels of annotation are sufficient to capture the most central structures necessary to build the knowledge base and whether the technique will prove sufficient to deal with a dynamically expanding corpus.

Bibliography

Bateman, J., Kasper, R., Moore, J., Whitney, R. (1989). A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model. Technical Report. ISI. Marina del Rey, Cal.

Bateman, J. (1990). "Upper Modelling: A general organization of knowledge for natural language processing." In: Proceedings of the International Language Generation Workshop. Pittsburgh 1990.

Bateman, J. (1992). The Theoretical Status of Ontologies in Natural Language Processing. Proceedings of the Workshop on ‘Text Representation and Domain Modelling – Ideas from Linguistics and AI’, TU Berlin, Oct. 9-11th, 1991. KIT Report 97.

Buitelaar, P. (1998). CoreLex. Systematic Polysemy and Underspecification. PhD thesis. Computer Science, Brandeis, University, Feb. 1998.

Church, K., Gale, W., Hanks, P. & Hindle, D. (1991). “Using statistics in lexical analysis.” In: Zernik, U. ed. Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale, NJ: L. Erlbaum. (pp. 115–164).

Fankhauser, P. & Teich, E. (2004). WordNet for lexical cohesion analysis. Proceedings of the 2nd Global WordNet Conference, Brno, January 2004.

Fellbaum, C. ed. (1998). WordNet. An Electronic Lexical Database. Cambridge, Mass., London, England: The MIT Press.

Firth, John Rupert. (1957). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis. Reprinted in: Palmer, F.R. ed. 1968, Selected Papers of J.R. Firth. Harlow: Longman.

Halliday, MA.K. & Hasan, R. (1976). Cohesion in English. Longman: London.

Halliday, MA.K. & Matthiessen, C. (1999). Construing Experience Through Meaning: A Language-based Approach. London, New York: Cassell.

Hindle, D. (1994). A parser for text corpora. In: Atkins, B.T.S. & A. Zampolli. eds. Computational Approaches to the Lexicon. Oxford: OUP. (pp. 103-151).

Hunston, S. & Francis G. (1999). Pattern Grammar. Amsterdam, Philadelphia: John Benjamins.

Levin, B. (1993). English Verb Classes and Alternations. A Preliminary Investigation. Chicago, London: The University of Chicago Press.

Sinclair, John. (1991). Corpus, Concordance, Collocation. Oxford: OUP.