Querying meta-data and object data -

Problems and elements of solutions in the domain of linguistic corpora

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
- Computerlinguistik -
Azenbergstr. 12
D 70174 Stuttgart
{heid, evert, berman}@ims.uni-stuttgart.de

Abstract

The purpose of this short presentation is to start some brainstorming about the requirements for query systems covering both meta-descriptions and object data descriptions. To do so, we first sketch a few assumptions about the nature of both kinds of descriptions, and about the kinds of queries intended. Thereafter we give concrete examples of the possibilities available in two research prototypes of corpus query systems, the IMS Corpus WorkBench, CWB, and the query tool of the MATE Workbench; a discussion of these to systems should highlight some of the main questions to be addressed in specifications for a query tool of the targeted kind.

Current text and dialogue corpora contain linguistic annotations; these are in fact classifications of single word forms or ``regions", i.e. sequences of word forms (chunks, sentences, utterances, turns, dialogues, etc.) from the point of view of different levels of linguistic description. For example, part-of-speech and lemma annotations are fairly standard for the European languages, and projects such as MATE(LE-4/8370, recently completed), have proposed annotations covering ``higher" levels, including, among others, syntactic chunks, coreference relations, dialogue acts, etc.

As well as annotations of object data, some corpora also contain descriptive meta-data:the language used, the social properties of the speakers, the utterance situations, etc. Such classifications (which have to be based on some commonly agreed on set of criteria) may concern ``regions" of different size: utterances or turns, but quite often whole dialogues, or texts, or recordings of some other kind.For many research purposes, it is necessary to be able to query sets of corpora, as well as individual corpora or documents, dialogues, etc. according to criteria related to both kinds of classifications. This can be done in different ways:

- The two kinds of classifications could be queried serially, by distinct query languages, in either of two orders: (i) First, a selection of documents to be queried linguistically is computed by means of an ``external" query, using meta-descriptions as search criteria; subsequently, those documents that satisfy the targeted meta-specification are queried ``internally", i.e. using the (linguistic) annotations as search criteria. (ii) First, a linguistic

query is performed; then, the result set is narrowed down according to meta-description criteria. This approach is generally inefficient, because the linguistic query has to be evaluated on the entire data collection.

- Alternatively, meta- and object data could be queried in parallel. That is, given a collection of documents (either seen as independent corpora or as subcorpora of one big corpus) annotated both at word level (with object descriptions, e.g. linguistic annotations) and at region level (with meta-descriptions and with some types of linguistic descriptions), a query would select material from the document collection according to constraints of both types. This approach requires a language capable of formulating both meta- and object data queries.

A precondition for any successful query according to meta-descriptions is the existence of agreed-on classifications. Anticipating the existence of such classifications, the following properties seem to be useful desiderata

- The values of meta-descriptions should be defined and organized in terms of type hierarchies.

- Different meta-descriptions may have different scopes within given corpora. For instance, the description of a situation may go hand in hand with a (putative) dialogue act annotation (and thus cover several turns of several speakers), while the description of the language spoken applies to the dialogue as a whole. This implies that the ``regions" to which certain meta-descriptions apply are contained in each other, perhaps even overlap (e.g. a turn of a dialogue and noise).

To gain some insight into the needs of query tools supporting a joint query of meta-descriptions and object data descriptions, we discuss the possibilities at hand in two different query systems, IMS CWB Corpus WorkBench[1], based on CQP language[2], and the MATE QueryTool embodied in the MATE Workbench[3] and based on the Q4M query language[4.]

CQP has been designed for the handling of large corpora of written text (currently in use with corpora of around 200 M wordforms). Object data annotations are so far limited to word forms, and meta-descriptions are supposed to be annotated only to regions. Multiple annotations of regions are possible, but hierarchical relationships cannot be represented. Current retrieval possibilites are optimized for object data annotations, but additions to the language specification are under way which would make annotations of word forms and annotations of regions queriable in the same way, and with the same mechanisms.

Q4M has been designed for the handling of massively annotated corpora (typically dialogue corpora), and on the assumption, that both individual word forms and ``regions" carry object data annotations, typically from different levels of description. In MATE, ``regions" could also be defined by means of their extension on the time line, which allows overlaps. Such regions can be organized in multiple parallel hierarchies. These properties could easily be used for meta-descriptions as well.

In the presentation, we will take a few examples from existing data (involving both meta-descriptions and linguistic object data descriptions), briefly discuss their representation in both corpus query systems and then present the queries needed (in the current state of the tools) to selectively retrieve linguistic material according to a combination of constraints of both kinds.

[1] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
[2] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/
[3] http://www.ltg.ed.ac.uk/~dmck/MateCode/
[4] http://www.ims.uni-stuttgart.de/projekte/mate/WB3/Q4M/001/docu/quer.html