Automatically assigned Syntacic Annotations

Gertjan van Noord <vannoord@let.rug.nl>

Table of Contents

Introduction

This document gives a short overview of the procedure which has been used to annotate the full D-Coi corpus syntactically, using the Alpino parser. A version of Alpino was used, which is for all practical purposes equivalent to the following version, which is the current version available for download at the Alpino website (at the time of writing this document):

```
Alpino (runtime), version 2007-01-30T17:24+0100
```

Procedure and parameters

The sentences which were parsed by Alpino are extracted from the D-Coi XML files using a simple style-sheet (available in the data directories). The tokenization decisions made in the D-Coi XML files (sentence splitting and token splitting) were all taken as is, and no attempts were made to correct errors (in contrast to the manually annotated sentences). The sentences in this format are available in the *Suites* subdirectory of the data directory.

The sentences were parsed by Alpino. For practical purposes, sentences were combined into sets of 2500 sentences, and for each set of 2500 sentences, an incarnation of Alpino was run for parsing. This took typically somewhere between 10 and 30 hours of CPU-time. In a few cases, the parser failed because it ran out of memory, or out of CPU-time, for a given sentence. In such cases, the parse of this sentence was aborted. No dependency structure is available for those sentences, but there is an XML file. These cases can be recognized using the *comments* element in the XML.

The relevant options that were used for Alpino are as follows:

```
-fast
user_max=600000
min_sentence_length=0
max_sentence_length=0
end_hook=xml
```

Evaluation

We can estimate the accuracy of the automatically assigned syntactic annotations by looking at the accuracy of the Alpino parser on the manually corrected D-Coi parses. For practical purposes, this experiment has been performed only on the syntactic annotations that were manually corrected in Groningen. To be more precise, the evaluation has been performed on the following sets:

```
WR-P-E-E   18-20
WR-P-E-H   9, 13, 20, 27, 36, 40, 47, 49, 50-52, 55
WR-P-E-I   1, 4, 6, 8, 11, 14
WR-P-P-B   1
WR-P-P-C   3-8, 11
WR-P-P-E   1, 2, 4
WR-P-P-F   6-8
WR-P-P-H   1-109
WR-P-P-J   1
WR-P-P-L   1, 3
```

We now give the results on these sets in the following table. Note that accuracy is expressed by means of the Concept Accuracy metric, which can be understood as the proportion of correct named dependencies. Formal definitions are provided in van Noord 2006 and references listed there. van Noord 2006 [http://www.let.rug.nl/~vannoord/papers/taln.pdf]

set	#sentences	CA%
WR-P-E-E	90	81.08
WR-P-E-H	2604	86.68
WR-P-E-I	240	87.55
WR-P-P-B	277	93.47
WR-P-P-C	937	89.42
WR-P-P-E	307	88.35
WR-P-P-F	397	85.47
WR-P-P-H	2267	90.70
WR-P-P-J	765	85.44
WR-P-P-L	1119	88.71
TOTAL	9003	88.28

There is one aspect that should be taken into account when we use these numbers to estimate the accuracy of the fully automatically assigned syntactic annotations. In some cases, the tokenization that was assigned automatically has been adapted for sentences that were part of the sets for which the syntactic annotations have been manually corrected. This implies that the figures above will typically be a somewhat optimistic estimate.

Representation

For each sentence, the dependency structure assigned by the parser has been saved in an XML file. For reasons of storage, the 2500 XML files for each set of sentences are packed together in two compact forms. The first form is simply a gzipped tar archive which can be unpacked using the UNIX tar command. In the second form, 2500 XML files are concatenated and compressed into a single file with extension *.dz* A corresponding *.index* file is created to be able to extract XML files from this compressed set of files, without the need to decompress the whole set. The Alpino Treebank tools understand this representation (inspired by *dictzip*), and documentation can be found in the User Guide of the Alpino Treebank Tools

[http://www.let.rug.nl/~vannoord/alp/Alpino/TreebankTools.html].