Analyzing German Noun Compounds using a Web-Scale Dataset – Report Week 3

TECHNISCHE UNIVERSITÄT DARMSTADT

UIMA Software Project WS 2010/2011
Jens Haase

1 Work done in the last week

1.1 Improving dictionary

The first thing of this week was to improve the igerman98 dictionary. For that the flag of each word is handled by the dictionary reader. The flag identify rules. These rules come in a seperated text file with the ending `aff` for affix. The rules mostly represent prefixes and suffixes, which change the word at the beginning or the end. All flags can also be combined. But this is currently not implemented.

The `Finder` class can now also be used as dictionary. It implements now the `IDictionary` interface. Using this dictionary with the current implementaion of the splitting algorithm is very unperformand. I tested the performace with a the `FinderPerformaceTest` class in the `test` dictionary. Most of the operation on the index are very fast. But sometimes it takes up to two seconds. Running the evaluation takes very long because the current splitting algorithm makes heavy use of the dictionary. After around 50 words of the corpus and 1 hour of waiting I stopped the program.

The first 50 result also showed me that the unigrams of the web1t corpus do not have a high quality. The corpus contains a lot of words that are not real words.

1.2 Improving left to right algorithm

With the improved dictionary I hope that also the recall of around 0.45 will be better. But in the first run the recall becomes more worse. It only was by around 0.43. Looking at the result, a lot words look very good, but never exactly like the correct one. The most error cases are the morphemes (position of brackets in textual representation).

In the next evaluation steps I ignored morphemes and only looked if the split was a the right position. That results in a recall of 0.826, which is very good compared to 0.43. This indicates that the real problem is not the split position. It is a problem of how to set the morphemes. Imagine the correct form of the word *Aktionsplan* is *Akt+ions+plan*, but the current algorithm only returns *Akt+ion(s)+plan*. For this example it could be a good idea to iterate over all words and check if the word combined with the morpheme is also a valid word (e.g. check for *ions*). All combination of these new words are candidates for the right split. For evaluation of the improvment see the next section.

1.3 Current Evaluation

With the current status of the algorithm the recall of correct splits is by 0.44. Without looking at morphemes we have a recall of 0.96. This mean 96% of all word are splitted at the right position. The words only differ in the position of morphemes.

1.4 Time Tracking

Date	Task	Needed time	Planned time
2010-11-25	Better dictionary reader	3	4
2010-11-27	Changed Evaluation (improve Splitter)	2	3
2010-11-30	Used Finder as dictionary (improve Splitter)	2	2
2010-12-01	Improving left-to-right algorithm, Report writing	6	4+2

2 Plan for next week

In the next week I will go more in detail with the morphemes and try to improve the recall value. I also want to try a new data driven algorithm as mentioned by Larson et al. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.35.7240\&rep=rep1\&type=pdf`. At the end I will try to compare the two algorithm.

Date	Task	Planned time
2010-12-02	Improve splitting algorithm	5
2010-12-04	New data driven algorithm	2
2010-12-06	New data driven algorithm	2
2010-12-09	New data driven algorithm, Evaluation, Report Writing	6