Analyzing German Noun Compounds using a Web-Scale Dataset – Report Week 7

TECHNISCHE
UNIVERSITÄT
DARMSTADT

UIMA Software Project WS 2010/2011
Jens Haase

1 Work done in the last week

1.1 Internal search cache

The first attempt to improve search speed is a simple least recently used (LRU) cache. This cache maps the last thousand search queries to the search results of the index. Queries that has been send before must not be searched again in the index. The cache improves the speed for the first 1000 splits from 30 minutes to 20 minutes.

1.2 Multiple indexes

The next step was to split the single index in multiple indexes, with the hope that the search on all these indexes in parallel will be faster than the search on one single index. But this was not true. I created two new indexes. One with nine single indexes and one with 26 indexes. Both were slower than the single index. As see above the splitting of the first 1000 words takes 20 minutes. The same task with a nine-splitted index takes 30 minutes and with the 26-splitted index runs over 50 minutes.

1.3 Improvements

All software improvements had nearly no effect on the search speed. Reading more about lucene optimizing leads to the fact that only other hardware like flash-based solid state disk can help to increase the speed.

A last attempted can be to create a smaller index. Since the splitting algorithm contains only word that are part of the dictionary, only n-grams that contain one of these words are interesting.

1.4 Time Tracking

Date	Task	Needed time	Planned time
2011-01-03	Implementing caching and testing results	2	0
2011-01-04	Build new indexes and testing result	7	6
2011-01-05	Test results and Report Writing	4	4

2 Plan for next week

In the next week I want to create a smaller index as mentioned above. Also the next ranking algorithm should be implemented.

Date	Task	Planned time
2011-01-06	Smaller index	4
2011-01-07	Ranking algorithm	3
2011-01-18	Ranking algorithm	3
2011-01-19	Evaluation and Report Writing	3