Analyzing German Noun Compounds using a Web-Scale Dataset – Report Week 4

TECHNISCHE UNIVERSITÄT DARMSTADT

UIMA Software Project WS 2010/2011
Jens Haase

1 Work done in the last week

1.1 Refactoring and Improving left to right algorithm

The first step of this week was to refactor the left to right algorithm. The goal was that the algorithm should return a tree instead of list. The tree structure has the benefit that it can be visualized. The visualization of the splitting algorithm can be used to see how the algorithm works and which errors occur. The new structure can later also be used for the ranking algorithm.

The old algorithm was moved to a separated package and marked as deprecated. It is only available to compared the result with the new one.

1.2 New Data driven algorithm

Next to the *left to right algorithm* there is now a new splitting algorithm. This splitting algorithm tries to split the word based on the amount of words that begin and end with the same letters as the word that should be splitted. A detailed description can be found at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.35.7240&rep=rep1&type=pdf`.

For the algorithm a trie implementation is needed. This implementation can be found in the `trie` package. The trie was only tested with the IGerman98 dictionary. Dictionaries with a lot of more words can be ran out of memory. For that a trie with a higher performance is required.

The implementation of the algorithm itself also returns a tree, as mention above in the new left to right algorithm.

1.3 Current Evaluation

Due to a bug in the evaluation class, the results of the last week are not correct. See the following table for the current evaluation of the algorithm:

	Left-to-Right (Old)	**Left-to-Right (New)**	**Data-Driven-Algorithm**
Recall with morphemes	0.44	0.47	0.16
Recall without morphemes	0.83	0.88	0.41

As we can see the new *left to right algorithm* works best. But it requires some future work to find the right morphemes. The splits are in most cases a the right position.

The new data driven algorithm returns very bad results. The recall without morphemes is more worse the the recall with morphemes of the *left to right algorithm*. I think improving this algorithm will never beat the *left to right algorithm*.

1.4 Time Tracking

Date	Task	Needed time	Planned time
2010-12-02	New Data Driven Algorithm (Trie)	4	4
2010-12-06	New Data Driven Algorithm	3	4
2010-12-07	Refactor/Improve left to right algorithm	5	5
2010-12-08	Report writing	2	2

2 Plan for next week

The focus for the next week is to improve the new left to right algorithm. Especially, the recall with morphemes should be higher. In the next week I also want to start to think about a ranking function. This should result in different concepts that can be implemented in the following weeks.

Date	Task	Planned time
2010-12-09	Improve splitting algorithm	4
2010-12-12	Ranking functions	3
2010-12-13	Ranking functions	2
2010-12-15	Report Writing	2