Analyzing German Noun Compounds using a Web-Scale Dataset – Report Week 8

TECHNISCHE UNIVERSITÄT DARMSTADT

UIMA Software Project WS 2010/2011
Jens Haase

1 Work done in the last week

1.1 Smaller index

To create a smaller web1t index only n-grams that contain dictionary words are added to the index. This reduces the index size from 23.4 GB to 12.8 GB. The execution time for the first 1000 stays at the same level as before (20 minutes).

1.2 Mutual Information ranking

As mentioned in week 5 the mutual information ranking algorithm was implemented. This algorithm checks the probability that two words appear in one n-gram. A problem of this algorithm is that the original word (not splitted) can not be ranked. That means this algorithm can not decide if a word should be splitted or not.

1.3 Current evaluation of ranking algorithm

The following table shows the current evaluation of the three splitting algorithms. Only the first 1000 words are evaluated to save time. For each algorithm we check if the correct word is at the first, second or third position. The Correct@1 value show how many correct splits are at the first position of the ranking. The Correct@2 value says if at the first or the second position is a correct split. The value in the braked do not check if the morphemes are set correctly, only if the splits are at the correct position.

Algorithm	**Correct@1**	**Correct@2**	**Correct@3**
Frequency Based	0.511 (0.698)	0.664 (0.778)	0.618 (0.657)
Probability Based	0.175 (0.243)	0.588 (0.751)	0.607 (0.63)
Mutual Information Based	0.419 (0.603)	0.476 (0.605)	0.415 (0.452)

1.4 Time Tracking

Date	**Task**	**Needed time**	**Planned time**
2011-01-18	Smaller index and mutual information algorithm	8	10
2011-01-19	Evaluation and Report Writing	3	3

2 Plan for next week

While the probability based method focuses more on words with less splits the mutual information algorithm focus on words with more splits. Bringing all this algorithm together can improve the results of the complete algorithms. Also a tree based ranking algorithm can result in better results.

The next week should be the last week with a focus on the ranking algorithm. In the week after that a UIMA component should be implement at the code should be cleaned up.

Date	Task	Planned time
2011-01-20	Tree based method	3
2011-01-21	Tree based method	3
2011-01-22	Combine algorithms	3
2011-01-26	Evaluation and Report Writing	3