NLP Assignment 1

Timur Yakupov

February 2017

1 Introduction

The task was to implement a Simplified Lesk. The Python programming language is used for my implementation.

2 Implementation

My implementation of Simplified Lesk is based on the description in one of the lectures (http://www.derczynski.com/sheffield/teaching/inno/3b.pdf).

There are a lot of methods:

- test(instances, context_size) - method that evaluates Lesk and caclulates accuracy.
- simplified_lesk(word, sentence) - Simplified Lesk algorithm itself
- get_all_senses(word, part_of_speech) - gets all possible senses of the word
- get_wordnet_pos(tag) - maps tags to WordNet part of speech names
- get_word_tag(word, context) - method for getting a tag of a word within the context
- ...

I decided to limit the context of the word when I take it from Senseval. So the test method in WSD.py file has context_size parameter which can be used to limit the context. For example, if you make it 5, the context will maximum be 10 words (5 from each side).

In simple words this is how my algorithm finds the sense of a word:

- Limit the context
- Tag the word and its context to find word's part of speech.
- take all senses for given word and its part of speech
- remove stopwords

- compute overlap of context and example sentences of all taken senses, the sense with the max overlap is the best sense

Sometimes due to tagging error there are no senses found. In that case all possible senses for the word are taken. The default sense is the first sense of wn.synsets(...) result. In case of Zero-Overlap the default sense counts as best sense.

The part with finding part of speech is actually the extension I made for my Simplified Lesk. It is made to ensure that the sense at least matches the part of speech.

3 Results

The algorithm was tested over the 300 Senseval example sentences. Here are some results in percentages:

When context is 7:

```
accuracy 0.74

Process finished with exit code 0
```

When context is 2:

```
accuracy 0.9466666666666667

Process finished with exit code 0
```

When context is 2 the accuracy is unbelievable big. This may happen because the context is too small and no overlap happens. But the default sense for word "hard" is exact as in the first Senseval sentences.

In fact, the bigger context causes lesser accuracy. I suppose that it happens because more context causes more overlap with words that does not actually refer to the word which sense we want to find.

4 Conclusion

The Simplified Lesk I implemented is far from perfection. It has low accuracy, but anyway it was a good experience.