Class project: Complex Word Identification

Pierre Finnimore

Abstract

1 Introduction

Initial words: 166

What is the task and why is it important?

The task is as follows: given a target word (or set of words) within a sentence, identify if the target is "complex". The data given for this task is a set of labelled sentences with targets. The labels were derived from a survey of both Native and Non-native speakers of two languages: English and Spanish.

We are interested in identifying word complexity for several reasons. Automatic extraction of complex terms could help with automated tutoring systems, Natural Language Generation, writing editing software, studies into second-language acquisition, political speech analysis, Machine Translation, as well as linguistic or psychological studies into the what people find complex.

2 Baseline system description

System descriptions in enough detail for the reader to be able to understand how to reimplement your baseline models and to appreciate why they are suitable for the task at hand.

3 Improved system motivation and description

4 Experiments on development set

Does your idea work as expected? Evaluate on the test set the baseline and the improved system, is it still the case? Identify examples in development data which help showcase why the improved system works better.

5 Learning curves

Plot learning curves for the trainable systems you experiment with. Are some systems better than others when less training data is available?

6 Examples of failed predictions

Identify examples where your improved system fails to predict correctly and propose ideas for future work to address them.

7 Conclusions

what have we learnt from your experiments that could inform future work