Indian Institute of Technology, Varanasi

B.Tech Project

Part - III

Word Complexity Identification

Submitted By:	Supervisor:	Co-Supervisor:
Prashant Goyal	Dr. A.K.Singh	Dr. P.Ahuja

April 21, 2018

Content

Abstract

Natural Language Processing is one of the most disruptive components of current AI technology and being able to interact with our devices in natural, 'human' way opens the doors to solve a huge variety of problems. Natural Language Systems requires a big corpus. The problem with big corpus is that it contains many complex words for certain target audience which unnecessarily increase the complexity of the system. It has been shown that the introduction of Lexical Simplification Module at beginning of pipeline outperforms other systems. So, I have tried to make such a system namely CWI.

Introduction

Text simplification systems aim to facilitate reading comprehension to different target readerships such as foreign language learners, native speakers with low literacy levels or various kinds of reading impairments. Identifying which words are considered difficult for a given target population is an important step for building better performing lexical simplification systems. This step is known as complex word identification (CWI).

The goal of the Complex Word Identification task is to provide a framework for the simplification for the first step in a Lexical Analysis pipeline. This is a simple, well-defined and yet challenging task that many in the community of Lexical and Text Simplification, as well as newcomers.

Dataset Description

The dataset used for developing the CWI module is downloaded from Shared_Task competition. The dataset consists of mixture of professionally written news, non-professionally written news (WikiNews), and Wikipedia articles. The number of instances for each training, development and test set is: 27,299 for training, 3,328 for validation and 4,252 for test.

Each sentence in the dataset was annotated by 10 native and 10 non-native speakers. Annotators were provided with the surrounding context of each sentence, i.e. a paragraph and then asked to mark words they think would be difficult to understand for children, non-native speakers, and people with language disabilities.

Format of Training Dataset

```
<ID> Both China and the Philippines flexed their muscles on Wednesday. 31 51
flexed their muscles 10 10 3 2 1 0.25
<ID> Both China and the Philippines flexed their muscles on Wednesday. 31 37
flexed 10 10 2 6 1 0.4
```

Each line represents a sentence with one complex word annotation and relevant information, each separated by a TAB character.
The first column shows the ID of the sentence.

Second column shows actual sentence where there exists a complex phrase annotation.

Third and fourth columns display the start and end offsets of target word in sentence.

Fifth column represents the target word.

Sixth and seventh columns show the number of native annotators and the number of non-native annotators who saw the sentence.

Eighth and ninth columns show the number of native annotators and the number of non-native annotators who marked the target word as difficult.

Tenth and eleventh columns show the gold-standard label for the binary and probabilistic classification tasks.

The labels in the binary classification task were assigned in the following manner:
0: simple word (none of the annotators marked the word as difficult)

1: complex word (at least one annotator marked the word as difficult)

The labels in the probabilistic classification task were assigned as “the numbers of the annotators who marked the word as difficult” / “the total numbers of annotators”. But task is to predict the probability which suited the problem the most.

Task Description

The task done in UnderGraduate project is a shared task competition task in which users were given multilingual dataset (German, English, French). Each team had to pick one language and perform the task of “Complexity of Word Identification” on that language dataset. Each team had to either predict whether the word is hard for target speakers/audience and what is the probability associated with it. The measurement method for accuracy score was not described earlier but they provided the validation data to tune hyperparameter.

Furthur task in this project is basically to suggest some synonyms of a target word. For a specific word a list of substitutes will be formed and top n words will be chosen and each of them will be ranked according to some “score” and will be replaced with the original word.

The substitute word must be chosen in the way so that the sentence does not lose its semantic meaning and also the syntactic structure.

Literature Review

1. Simplifying Lexical Simplification: Do We Need Simplified Corpora?

This paper tries to find the complex the word in a regular corpora and replace with the most suitable synonym of that word. The authors introduced LIGHT-LS, a novel unsupervised approach to lexical simplification which does not rely on the rule based word substitution for every complex word, which makes it applicable in settings where such resources are not available. With the state-of-the-art word vector representations (GloVe vectors) at its core, LIGHT-LS requires nothing but a large regular corpus to perform lexical simplifications.

They represented every word in a GloVe vector form of 200 dimensions and applied cosine similarity for finding the top 10 words similar to the target word and then employed Berkeley Language Modal (Pauls and Klein, 2011) to compute the likelihood of those 10 words to fit into the sequence using the bigrams and trigrams.

2. Unsupervised Lexical Simplification for Non-Native Speakers

In this paper authors find a way to simplify a text using an unsupervised lexical simplification approach applied on a dataset of subtitles extracted from IMDB. They try to find a way to identify a complex word in a sentence and generates its possible substitutions and choose a subset of them, rank them and replace it.

We have also introduced NNSeval, a new dataset for the evaluation of LS systems which targets the simplification needs of non-native English speakers

They proposed 2 constraints on the substitute word, given the target word, its POS tag and context aware embedding representation, the substitute word follows 2 rules/constraints:

- ❑ The word must share the same POS tag as the target word.
- ❑ The word must not be a morphological variant of the target word.

Methods/Models applied

The methods I tried do not use any machine learning or Natural Language Processing for that matter. I simply tried some statistical analysis and hyperparameter tuning. All these methods are based on how many annotators marked a specific target difficult. The 3 models I used are described below.

1. Baseline Model:

This model was based on a very simple approach, here I considered a word difficult if it is considered difficult by even a single annotator of any kind and the length of the phrase/word is larger than 10 characters. The accuracy measured on the training dataset itself was 76.93 and on the validation dataset was found to be around 68.82.

2. Restricted Baseline Model:

In this model I used a slightly more restrictive approach in training where if a phrase/word is marked difficult by even a single native-annotator and by at least single non-native annotator, it was predicted as difficult. The accuracy measured on the training dataset itself was 72.46 and on the validation dataset was found to be around 66.12.

3. Context Based Model:

This model uses a modified version of the above two approaches. Here I used a context window of size 2. Then if it was marked difficult by any annotator then it was assumed difficult. The accuracy measure on training dataset was 84.45 while on validation dataset it was just 35.12. The result can be explained by the fact that training algorithm stored all cases and didn't actually learn anything and most of the examples in validation dataset were not found in training dataset.

Applications and scope in chemical engineering

It has been shown that individuals with low-literacy levels or who suffer from certain clinical conditions, such as Dyslexia, Aphasia (Devlin and Tait 1998) and some forms of Autism, often hindering them incapable of recognising and/or understanding the meaning of texts. Impairments that cause the narrowing of one's vocabulary can be severely crippling: the results obtained by (Hirsh, Nation, and others 1992) show that one must be familiar with at least 95% of the vocabulary of a text in order to understand it. Lexical Simplification (LS) and CWI specifically aims to address this problem by replacing words that may challenge a certain target audience with simpler alternatives.

In chemical engineering and other type of sciences, often research papers are filled with very complex terminologies hindering the reader to completely focus on main research described in the text. Thus Lexical Simplification and CWI might help to reduce the complexity of the paper and enhance the readability.

Vision

The accuracy on validation set by the returned by above applied models is very low as they are very naïve and do not use any morphology of the sentences and phrases. They also not apply much statistical analysis and Natural Language Processing technique to give any useful information about the data.

In the next phase of the project, I plan to use some more NLP like parsing, tagging and NER and some Deep Learning to enhance the accuracy of the system.

References

- https://sites.google.com/view/cwisharedtask2018/home
- http://alt.qcri.org/semeval2016/task11/
- Butnaru, Andrei & Ionescu, Radu Tudor. (2018). UnibucKernel: A kernel-based learning method for complex word identification.
- Glavaš, Goran & Stajner, Sanja. (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora?. 10.3115/v1/P15-2011.
- Paetzold, Gustavo. H & Specia, Lucia. (2016). Unsupervised Lexical Simplification for Non-Native Speakers