CS 388: Natural Language Processing Spring 2016

Project Report — Original Language Detection

Students: M. Denend, J. Hoffmann, L.Prakash *Prof. Ray Mooney*

1 Project description

(a) Description

The goal of this project is to be able to find the original language a text was written in, from the translated English version. We believe that even with completely bilingual translators, languages have a specific structure that will influence the translator in the way he creates the translation. We hope to recover this structure during our task.

(b) Methodology

From Project Gutenberg, we built a dataset of books written in different languages, and translated into English. We then cut these books into slices of 500 sentences, and compute features on these slices. After feature extraction, we feed these features into a classifier and test it on a completely different set of books.

2 Building the dataset

(a) Pitfalls

Since we built it ourselves, our dataset is rather small. We therefore were extremely careful not to overfit on books particularities that are not specific to the language, such as topic, style, author, geographic places, time period. We believe the way we've constructed the dataset satisfies these criteria.

(b) Description of the dataset

When building the dataset, we used books written from 1700 to 1900 to control the time period as best as we can. We made sure that we never used the same author nor translator (?) in train and test, and we tried varying the topics between train and test as much as we could once the previous criteria were met. When using Lexical Features, we only remembered the most common 1000 English words to control overfitting on topics as well. The dataset we used is further described on this table:

[insert TABLE]

3 Features

(a) General Description

When computing the features, we cut the books in slices of n sentences. With $n = 1$, we get a lot of training examples, but finding out the original language from only 1 sentence seems unreasonable (which was proven

by experiment). With the whole book as a sample, we didn't get enough training instances. We chose $n = 500$, as it provided good balance between number of train samples, and informativeness of the samples.

(b) Lexical Feature

Three different kind of lexical features were used:

1000 most common English words After overfitting on words like "Moscow" or "Madame" when using unigrams, we decided to only use the 1000 most common words in English to avoid overfitting on topics, local idioms, or geographic places.

POS unigram We only used POS unigrams, since we believed using bigram or trigram would just add redundancy with the Parse features.

Etymology For each word used, we computed the etymology of the word from parsing Wiktionary [2], and added all possible etymology as a feature. Rational being that if you're translating from a latin language, you may be more likely to use a latin word when choosing between synonyms. We only included the etymologies of nouns, verb, adjective and adverb, since we believed preposition, pronouns and others wouldn't be significant and would only drown the relevant data.

(c) Parse Features

Inspired by [3], we ran the Stanford parser on our slices, and computed the parse trees. By running a BFS, we then compute the generation rules and count them. We don't take into account the leaves of the trees to avoid redundancy with lexical features (and possible topic overfitting).

(d) Homemade Features

We also computed all the features of [1], which include length of sentences, length of words, ratio of pronouns, nouns, complex sentences etc. [MATT ADD WHATEVER YOU WANT].

4 Classification task

(a) Features Extraction

(b) Comparisons of different algorithm

5 Experiments

(a) Description of different experiments

We conducted the following experiments:

- using unigrams (overfitting on topics)
- using the same books for train and test
- using the same author for train and test

On top of that, we also inquired the effect of different features on the dataset:

- Most common words only
- Most common + etymology
- POS only
- All lexical features (Most common words + etymology + POS)
- Homemade only
- Parse only
- Most common + etymology + Homemade
- All features

(b) Results

[insert TABLE]

6 Conclusion

References

[1] Gerard Lynch and Carl Vogel. Towards the automatic detection of the source language of a literary translation. 2012.

[2] Christian M. Meyer and Iryna Gurevych. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford: Oxford University Press, November 2012.

[3] Sze-Meng Jojo Wong and Mark Dras. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610. Association for Computational Linguistics, 2011.