CS 388: Natural Language Processing Spring 2016

Problem Set 4 — April 21, 2016

Prof. Ray Mooney, TA Swadhin Pradhan *Student: Jessica Hoffmann*

1 Our project and this paper

Our project:
Our project is about Original Language Identification. From a text in English translated by a human from another language, our goal is to identify the original language of the text.

Chosen paper:
The paper I've chosen is *Exploiting Parse Structures for Native Language Identification* by Wong and Dras (2011) [5]. If Native Language Identification (abbreviated NLI is the rest of the text) is not exactly the same task as Original Language Identification, the tasks share some strong similarities. We expect to transfer some of the technics from this paper to our task.

2 Paper summary

One sentence summary

By including parse features, this paper reduces the error in NLI up to 30% compared to using lexical features only. NLI has some useful applications, including terrorist threat identification (when a message is intercepted, we may learn about the country of origin of the threat).

Data

The data is the *International Corpus of Learner English* (ICLE), compiled for the task of NLI. All the writers are supposed to have the same English proficiency, and to be in the same age range (it's a classical corpus in NLI). The dataset they use contains 7 languages, with 95 essays by language. After splitting, they get 17,718 train sentences and 6,791 test sentences.

Features

Lexical features
Based on previous studies, they've established that the best set of lexical features they can use is: A list of function words empirically established in [4] (binary feature), character bi-grams (best performing n-grams) (frequency feature), and the most frequently occurring PoS bi-grams and tri-grams (binary feature).

Parse features Parse features used include horizontal slices of the parse trees, presence of a subset of features (chosen through feature selection (IG)), and cross-sections for the RERANKING model. Parse rules are not (or very little) lexicalized to avoid overfitting on the topics.

Algorithms

After ruling out an SVM classifier, they use a MaxEnt machine learner.

3 Relation to our project

(a) Limitations

We believe this paper uses the parse features in a very restricted way. In particular, we want to explore using non-binary features (e.g. frequency of parse rule instead of just presence).
Contrary to their task, literary translations won't be prone to errors. A big part of NLI is identifying which native language will lead to which type of error, which is completely irrelevant in our task.
Their task is also limited in size, since they're using a classical corpus, when we can use any book in the public domain that has an English translation.

(b) Extensions

Since we're not limited to a single corpus, we will use a large amount of books (or chapters from books), using cross-corpus evaluations in the manner of [2] to make sure we're not overfitting on the author's style.
We will keep the lexical features mentioned in the paper, and maybe add the rare features described in [1].
We also want to add "homemade" lexical features based on the etymology of the words used in the case of synonyms (for instance, you may be more likely to use *damnation* rather than *doom* if you're translating from a latin language, and inversely if you're translating from a germanic language).
We want to improve the parse features, by making them not binary, but also by incorporating vertical features.
We want to also add some general features, including size of sentences, frequency of pronouns, active/passive voice etc. We will also try a larger set of algorithms, following the lead of [3] including Random Forest, SVM, and Gradient Boosting methods, or an ensemble of them, which are well-known to perform well on classification tasks.

4 Conclusion

We wish to transfer NLI features to our task, Original Language Identification. We will keep most of the lexical and parse features described in this paper, since they are relevant to our task, but will switch to a different dataset that we will create ourself, add some intuitive features for our problem, and use different machine learning algorithm that have been shown to work better with classification tasks.

References

[1] Julian Brooke and Graeme Hirst. Measuring interlanguage: Native language identification with l1-influence metrics. In *LREC*, pages 779–784, 2012.

[2] Julian Brooke and Graeme Hirst. Robust, lexicalized native language identification. 2012.

[3] Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57. Citeseer, 2013.

[4] Sze-Meng Jojo Wong and Mark Dras. Contrastive analysis and native language identification. *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–61, 2009.

[5] 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.