Matthew Denend
NLP HW 4

Reviewing Towards the automatic detection of the source language of a literary translation, by Gerard Lynch and Carl Vogel [2]

Problem Description: Lynch and Vogel [2] try to figure out the source language of translated literature using 19th century novels available to the public domain. They look at English, Russian, German, and French as their languages, and see if they can accurately determine the source language, and they try to determine which features are best for determining the source language with their corpus.

Motivation: Lynch and Vogel do not explicitly mention a motivation, but there are a couple of motivations for figuring out the source language for translation using machine learning:
1. When something gets translated, it may end up being a bit more awkward or ungrammatical than the author who wrote the document in the original language intended (called translationese [4]). Being able to detect this may help others better determine the original source language of text in case a work has an unknown source language.
2. In one of the noted papers in the previous research section, the authors cited a paper done by Baroni and Bernardini done in 2006 [1] that compared machine learning testing accuracy to humans differentiating between whether an article in the Italian publication *Limes* was written in its original language or was translated from another language. Only one out of ten humans outperformed the machine learning methodology. Hence, the paper suggests that machine learning can outperform most humans in identifying whether or not text was translated.

Dataset Used: Lynch and Vogel [2] took 20 different classic novels with 4 different languages: English, French, Russian, and German. They used 5 different novels for each language. No two books were written by the same author, and no two of the same translators were used either, in order to ensure diversity. All books were written in the 19th century to prevent any issue from arising with copyright law for books published in the United States. They also selected novels that were at least 200 KB in text size... so all novels/novellas were of good size. They took a 200 KB chunk of each of these texts and divided them into twenty 10 KB chunks per text, giving a total of 400 different chunks across the 20 novels. 18 of the 20 chunks per book were used in training and the other 2 were for testing.

Methodology: Lynch and Vogel [2] used the Weka Java toolkit for machine learning, and the TreeTagger POS tagger for POS parsing. The authors used three different types of classifiers: Naive Bayes, SVM, and Simple Logistic. For each of these classifiers, they ran two different types of tests: one in which they used the same 18 chunks for training and 2 chunks for testing for each book, and the other type was ten-fold cross validation, meaning that they rotated the testing chunks to ensure that all chunks in the book were tested against the training of the remaining chunks. Ex. If chunks 3-20 were trained and 1 and 2 were tested, then another test would have chunks 1-18 trained and 19 and 20 tested. These 10 runs were averaged to get one accuracy. The other dimension in the test was how many features were used: Just the 19-doc level tests (see paper [2] Table 2 for what these features are); taking the top 50 chi-squared performing features from a combination of 100 POS Bi-Grams, the 19 document features, and 15 top word unigrams; and also grabbing the top 30 ranked features from the same set.

Key Takeaways: Lynch and Vogel [2] provide several. Previous literature cited only used an original and translated language, this work was possibly one of the first to try four different languages (but never mentioned the approach being novel), and shows that it's possible to relatively accurately identify the source language of the translation. Also, some keys were actually pointed out in several common words such as prepositions (toward in German), contractions (especially in Russian), and certain other words appearing more often in certain languages. This approach to detecting language translation for the most part is a relatively solid approach, and something we'll have to consider, but not without flaws.

Strengths and Limitations: Expanding from 2 to 4 languages is a nice step [2], but what about expanding to more languages? Specifically, it would be interesting to see if adding more Indo-European languages (possibly Spanish, Greek, Italian, Hindi) would make identifying original source language harder.
The authors ensured that their data was diverse by not repeating the same translator or author twice, ensuring that they have a wide variety of data. Their methodology for the most part seemed reasonable, and using both three classifiers and three sets of features offered quite a great deal of flexibility. Still though, particularly with the features, I think running simple tests on seeing specifically which set of features give the best results independent of the others for both the POS bigrams and word unigrams (lexical features only) would have been interesting as well for comparsion, but they didn't do this. They also don't consider parse trees, which according to another paper [3], it's possible to reduce the error of native language identification up to 30% as opposed to using just lexical features. So adding parses on top of document-level features and POS bigrams and word unigrams should help.

This paper has one glaring limitation that we definitely wish to address: it uses the same training set and testing set! So, we need to introduce more books of the same source language translated into English from various languages that we've trained on, and see if we can identify the language of origin in a few cases: when the book is different but the author and translator are the same, when just the translator is the same but the author is different, when the author is the same but the translator different (if it's possible to find such novels), and when neither are the same, but the source language still is. Hence, this would steer us toward our main idea for our project: To see if we can build a parser that can identify the source language of a translated text using various author/translator combinations. Our intuition is that the authors were able to find unigrams and features that seemed to vary widely over the works, and the authors in [2] were able to pinpoint certain words; notably excluding proper nouns or words unique to a certain language, like Monsieur or Madame in French; that appeared much more commonly in certain languages. We think this will work because we think the grammatical differences and translationese should be detectable when translated into English, regardless of the translator or author. We think this should be enough of a difference across the languages to try changing the testing source and seeing how much results change. Furthermore, the suggestion of adding other languages to this current test, if this is possible, can give us an idea of whether we can fit more languages than just four. Whether or not this works will be a secondary step if we can get the different domain testing to work first, as we don't want to change too many variables in this experiment. We will be changing both the dataset (more books, possibly languages) and methodology (adding parse tree detection) slightly. We may keep some of the novels the same from the paper as a baseline, but we'll have to add more novels to give us possibly more interesting results.

Conclusion: All in all, I've analyzed this paper [2], and found most of the methodology to be worthwhile, so there's a ton of good from here that will go towards our project. This being said, there are some glaring limitations that we want to address in our project to see if we can come up with a more flexible methodology for identifying the source language of a translation by adding parse trees as a detection methodology, and trying to use different books from the same source language and three combinations of authors/translators; where both the author and translator are the same as a book in our corpus, where the translator remains the same as a book in our corpus, and where both the author and translator differ but the source language is the same as books in our corpus, and possibly by adding more Indo-Eurpoean languages.

References:
[1] Marco Baroni and Silvia Bernardini. . A new approach to the study of translationese: Machine-learning the difference between original and translated text. Literary and Linguistic Computing, Volume 21 Issue 3, pages 259-274. Oxford University, 2006.
[2] Gerard Lynch and Carl Vogel. Towards the automatic detection of the source language of a literary translation. *In Proceedings of COLING 2012*, pages 775-784. International Committee on Computational Linguistics, 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.
[4] https://en.wiktionary.org/wiki/translationese Definition of translationese. Last Modified 1/19/16.