CS 388 - Natural Language Processing - Project Proposal
Bradley Beth & Nathan Clement

1 Project Description

We would like to investigate application of statistical NLP techniques to Native Language Identification (NLI). In particular, can we construct a system such that given a text α produced by a speaker β in English, can we deduce the native language (L1) of β? As a simple binary classification problem, β = *{English | ¬English}*. Beyond that, can we extend classification to multiple languages as categories (either on a one-to-one, language-to-category basis or language family (with variable granularity) as category (e.g., Romance, Germanic, Sino-Tibetan, etc.))? Additionally, what categories of α maximize efficacy of training?

2 Data Source(s)

Our primary data source will be the *1 million word* **Vienna-Oxford International Corpus of English (VOICE)** (http://www.univie.ac.at/voice/). It represents 50 different L1 language-backgrounds spread over 1,250 different speakers. Each of the files in the corpus contains L2 English speech. Using VOICE is a novel approach as it was actually compiled as a speech corpus rather than as a text corpus. We intend to use the XML-formatted *transcriptions* of the corpus as training data. This allows capture of 'para-linguistic' phenomena (e.g., laughter, pauses) as features as well as words. As each of the files records a conversational event, the data will contain (1) interaction among speakers, (2) different domains of discourse, and (3) the spontaneity of speech. Previous work in the area of NLI appears to focus on edited writing samples rather than extemporaneous speech.

Additionally, we may use the **International Corpus of Learner English, v.2 (ICLE2)** (http://www.uclouvain.be/en-277586.html) if time allows. The Linguistics Representative of the UT libraries has placed an acquisition request for the corpus. This *3.7 million word* corpus is composed of essays written by English-language learners with a variety of native languages (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish).

3 Potential Algorithms / Experiments

At the simplest level, we will use *n-grams* as a language model. Positive results are theoretically plausible due to language education research on the importance of phrasemes and collocations in developing L2 fluency.

As time allows, we may study the effects of:

- modeling syntactic structures through the use of PCFGs (annotation-dependent)
- diversity of vocabulary (both lexemes and phrasemes)
- various computational approaches to classification (e.g., Naïve Bayes, logistic regression, SVM, etc.)

4 Previous Work

NLI is of multi-disciplinary interest:

- Work in **NLP / CompLing** has centered around NLI for security-motivated classification of documents and as part of a larger profiling platform (alongside gender, age, ethnicity, etc.).
  - *Representative Authors*: Moshe Koppel (Bar-Ilan University, Israel), Graeme Hirst (University of Toronto), Mark Dras (Macquarie University, Australia). Not much from the U.S., although Pennebaker did co-author with Koppel on at least one related paper.
- Work in **Language Pedagogy** has concentrated on descriptive error analysis in L2 acquisition of English and prescriptive modeling for maximal fluency.
- General **Corpus Linguistics** work has focused on modeling English as *lingua franca* and modeling emerging interlanguages among various L2 English learners.
- **Speech Recognition** has explored phonetic and acoustic properties of L2 English speech, while largely ignoring syntactic/semantic structures. Transcriptions of speech corpora have been lightly studied for modeling spontaneous speech. We have not found any work at the point of intersection.