You're Not From 'Round Here, Are You? Naive Bayes Detection of Non-native Utterance Text

Laura Mayfield Tomokiyo and **Rosie Jones**
Language Technologies Institute
Carnegie Mellon University
{laura.tomokiyo,rosie.jones}@cs.cmu.edu

Abstract

Native and non-native use of language differs, depending on the proficiency of the speaker, in clear and quantifiable ways. It has been shown that customizing the acoustic and language models of a natural language understanding system can significantly improve handling of non-native input; in order to make such a switch, however, the nativeness status of the user must be known. In this paper, we show that naive Bayes classification can be used to identify non-native utterances of English. The advantage of our method is that it relies on text, not on acoustic features, and can be used when the acoustic source is not available. We demonstrate that both read and spontaneous utterances can be classified with high accuracy, and that classification of errorful speech recognizer hypotheses is *more* accurate than classification of perfect transcriptions. We also characterize part-of-speech sequences that play a role in detecting non-native speech.

1 Introduction

Native speakers, while deeply individual in their use of language, share intuitions about the meaning and cadence of speech that come from many years of using it as the primary means of communication. Selection of the words that best express an idea in speech is often nearly instantaneous, yet is influenced by a complex collection of factors, including collocational conventions, rhythm, register, and semantic and pragmatic context. Variability notwithstanding, one might expect to find patterns in native speech that mark it as native.

For non-native speakers who are still learning a language, the production of meaningful speech is an entirely different task. In addition to the mechanical difficulties of pronouncing unfamiliar words, such obstacles as limited access to lexical terms and lack of intuition about the syntactic and semantic integrity of an utterance hinder non-native speakers in their efforts to speak.

The question that we seek to answer in this paper is whether native and non-native utterances of English are sufficiently distinct in character that they can be classified using statistical methods, namely naive Bayes classification.

Naive Bayes classifiers are often used to classify documents according to topic (Lewis, 1998). The classification of test documents is based on the probability of finding individual features (words) in the class based on the training set. Naive Bayes classifiers work well in combining the contributions of individual words across a large vocabulary, and also take into account the overall class probability. They compare favorably to other classification techniques when the class distributions are not radically skewed (Yang and Liu, 1999).

Machine learning techniques have also been used to categorize documents according to publication of origin, e.g., whether an article is likely to have appeared in *Time* or *Newsweek*. Source classification can be a more difficult problem than topic classification because the decision must be based on subtle differences in how words are used, not which words are most common. Content is determined by current affairs, which are common across the media in a given time-frame, whereas stylistic concerns are shaped by habits and preferences of the individual writers and editors. One might expect these kinds of effects to hold true for non-native speech detection, where the topics of speech are similar across speakers, but realization is different. These expectations have been formalized in such theories as transitional competence (Corder, 1967) and interlanguage (Tarone, 1978). It was shown in (Argamon-Engelson et al., 1998) that a document can be identified as coming from *Time*, *The Daily News*, *Newsweek*, *Times* editorials, or *Times* news using a decision-tree-based approach. Importantly, the features used here consisted of function words, which are commonly filtered out during feature selection for content-based classification. They also used part-of-speech tags as features in their classifier, suggesting that patterns of word use are more important in this task than the words themselves. Mosteller and Wallace (1984) showed that using Bayesian inference on function words can be effective for author identification.

Nativeness classification based on acoustic features has been used to detect foreign accents. Fung and Liu (1999) trained a hidden Markov model (HMM) to discriminate between native and Cantonese-accented English using energy and formant characteristics. Teixeira, Trancoso, and Serralheiro (1996) also used HMMs in a 6-way accent identification task, training a full set of accented acoustic models for each language pair, recognizing using each set of models in parallel, and choosing the model set with the highest acoustic score. Acoustically-derived classification, however, requires access to the acoustic features, which may not be readily available.

In this paper, we investigate classification of *text* as native or non-native. A text-derived classification result may be used as part of a feedback system in a speech recognizer: if the recognizer output fits a non-native profile, the utterance can be re-recognized with customized acoustic models or a customized lexicon. Text-based classification is appropriate if the recognition component is to be treated like a black box in a natural language understanding system; customized parsing and dialogue modeling can be invoked if the recognizer hypothesis indicates that the speaker is non-native. It may also be useful to consider nativeness in language modeling. Language model training text can be tagged as native or non-native when the model is built, and the role of the language model or models in the search can be modified if the initial hypothesis flags the speaker as non-native. Finally, text-based classification can be used to determine whether the author of a piece of text, such as an electronic mail message or a web page, is a native speaker; this information may contribute to improved parsing or information extraction.

The paper is organized as follows. In Section 2 we describe our target users and the speech data that was collected from them. The transcription protocol is also discussed here. In Section 3 we provide an overview of naive Bayes classification, the software that was used in our research, our text preparation and features used in the classifier. In Section 4 we describe our experimental methodology. Experimental results are presented in Section 5, with a discussion in Section 6.

2 Speech Data

In this section we detail the speech data used for our experiments. We describe the native and non-native speakers who were recorded, the recording conditions, and the types of transcriptions of their speech used in our experiments.

2.1 Speakers

It was our aim to choose speakers who must use English on a daily basis to communicate with colleagues and classmates but who experience significant difficulty in speaking it. These are potential users of natural language understanding systems who may be frustrated by the system's inability to adapt to non-native speech of their level.

45 speakers were recorded for this study. Of these, 31 were native speakers of Japanese, 8 were native speakers of English, and 6 were native speakers of Chinese. Non-native speakers were selected for length of time studying English in their home country (6 to 8 years), length of time in an English immersion environment (6 to 12 months), and self-reported comfort speaking English (3 on a scale of 5). Additionally, speakers were given a proficiency test (SPE, 1987) and a subset of 10 speakers with similar scores was chosen for controlled experiments.

2.2 Task and recording

All recording was done in a quiet room using a close-talking microphone. Speakers were alone in the room while recording. Both read and spontaneous speech were collected.

2.2.1 Read speech

For the read speech task, speakers read three articles of children's news; these selections were similar in content to common speech databases such as Wall Street Journal (LDC, 2000) but had a reduced vocabulary and lower difficulty that was better suited to our speakers. Of the three articles that each speaker read, one was read by all speakers, and the other two were read only by that speaker. The articles averaged 50 sentences, comprising around 1400 words, in length. The read speech task was performed by the native speakers and the proficiency-controlled subset of the Japanese speakers.

2.2.2 Spontaneous speech

All speakers performed a spontaneous task in the tourist domain. Speakers were prompted in their native language with sights and activities they might find interesting, and were instructed to ask questions of an agent about how to see or accomplish them. This is similar to the elicitation approach described in (Mayfield Tomokiyo and Burger, 1999). Conversations averaged 54 turns, with just over 1000 words per speaker.

2.3 Transcription

All recorded data was transcribed and verified by separate annotators. Transcription conventions followed those used by the Linguistic Data Consortium (LDC) in transcription of CALLHOME (LDC, 2000) with some extensions to support annotation of deviation from read text and non-native disfluencies.

2.4 Recognition

Recognizer hypotheses for the read and spontaneous speech recordings were produced by an HMM-based large vocabulary continuous speech recognition system (Finke et al., 1997). The baseline word error rate of this system on Broadcast News F0 data with trained newsreaders is 9.4%.

3 Naive Bayes Classification

3.1 Overview

A Naive Bayes classifier incorporates information about statistical priors on the target classes as well as the features present in each example. A test example is classified by assigning it to the class calculated as most likely to have produced it. For an utterance u which may assigned to a class c, its score is calculated as follows:

$$P(c|u) = \frac{P(c)P(u|c)}{P(u)}$$

When choosing between classes, we need not calculate the probability directly, since we want only to find the maximum score, and $P(u)$ is constant across classes.

$$argmax_{c_i} P(c_i|u) = argmax_{c_i} P(c_i)P(u|c_i)$$

Now, we have insufficient data to reliably estimate the probability of every possible utterance for arbitrary native and non-native speakers, since this set is infinite. Even simplifying to sequences of POS-tags will not permit completely reliable estimation of probabilities of utterances. Instead, we treat utterances as consisting of an unordered set of independently occurring features. Though this independence assumption may not be statistically true, it has been shown that for classification this kind of assumption does not substantially harm accuracy (Lewis, 1998; Domingos and Pazzani, 1997).

Thus for an utterance with features $f_1..f_n$ we can say

$$P(u|c_i) \approx \prod_{f_j \epsilon u} P(f_j|c_i) \tag{1}$$

leading to the selection of class c_i according to

$$argmax_{c_i} P(c_i) \prod_{f_j \epsilon u} P(f_j|c_i) \tag{2}$$

Our features consisted of all word pairs, in some cases word triples, and all of their constituent words. These word sequences will be referred to as unigrams, bigrams, and trigrams in this paper.

3.2 Text classification toolkit

The Rainbow statistical text classification package (McCallum, 1996) was used for all classification experiments. Rainbow implements a Naive Bayes classifier for text, with a number of features specialized for text applications.

3.3 Text Preparation

Since all documents were based on speech, no capitalization or punctuation was available. We treated spaces and the contraction-marking apostrophe as word separators, and included all unigrams and all bigrams as independent features. Trigrams were also used in some experiments. Calculation of probabilities was based on a multinomial event model, where feature probability is based on frequencies within documents, rather than just binary occurrence statistics.

3.4 Part-of-speech tagging

In some of the experiments we will describe, part-of-speech tags are used as input to the classifier instead of words. Part-of-speech tags were produced by the MXPOST toolkit (Ratnaparkhi, 1996). Bigrams and trigrams over the POS-tags were also used, as described above.

4 Experimental Design

The task of learning distinctive features of nativeness in text was framed as a document classification problem. That is, each article read by a speaker, or all of a speaker's side of each conversation, was treated as a unique document. Training data consisted of a set of documents labeled as native and set of documents labeled as non-native (for non-binary classification tasks documents were labeled with the speaker's native language). The classifier's task, then, was to build from the training documents a model of patterns distinguishing native and non-native language and use that model to make judgements about the nativeness of new documents. Training and test sets never contained the same speaker, and there was never more than one document from an individual speaker in a training or test set. Where training and test documents were from identical conditions (shared or unique, described below), we used 70% of documents for training, 30% for testing, with 20 random train-test splits, and with results averaged over the 20 runs.

4.1 Read speech

Although one might not expect read speech to vary from speaker to speaker, particularly when the text is the same, our preliminary investigations suggested that there are actually significant differences in the types of reading errors native and non-native speakers make that can be described with statistical measures such as perplexity and Kullback-Leibler divergence. Because our target speakers have had extensive exposure to written language but little experience with conversation, a read speech task represents a far lighter cognitive load than a spontaneous task. Errors in reading, then, might be considered more representative of fundamental differences in the linguistic models of native and non-native speakers, and as such strong bases for classifier training.

The proficiency-controlled subset of Japanese speakers was used in read speech classification. Articles from these speakers were contrasted with articles from eight native speakers.

As was described in Section 2.2, each speaker read three articles, one of which was common to all speakers, and two of which were read only by that speaker. Because the shared articles are all realized slightly differently, they provide a unique source of training and test data. There will be no distributional bias of lexical items, since the source text was all the same, so a classification decision will be based only on the character of reading errors.

While the shared articles are a valuable source of data, accurate classification of one article may not mean that the model will generalize to other articles. The unique articles read by each speaker are more representative of conventional document classification training and testing data, in which documents may share a topic or author or genre but are not transformations of a single base text.

With these considerations in mind, we defined four types of evaluation to determine in which situations detection of non-nativeness may be appropriate:

(A) train on shared article, test on shared article

(B) train on unique articles, test on unique articles

(C) train on shared article, test on unique article

(D) train on unique articles, test on shared articles

In (A), classification is based only on individual differences and is minimally affected by the base text. In (B), the model will be general, but it may be difficult to construct a useful model because differences in the base texts may overpower the subtle differences in native and non-native language, and we have less data available for building the model. In (C), a very specific model of reading errors in certain contexts will be built, but testing on different articles may show that the model does not generalize. Finally, in (D), a general model will be built, and successful classification of the shared article will show that the classifier is modeling nativeness, and not topic or lexicon, but may be open to concerns about external validity.

4.1.1 Working from transcripts

In the experiments described in this paper, we evaluated classification of both human-transcribed speech (transcripts) and the output of automatic speech recognition (recognizer hypotheses). This is a common practice in evaluations of systems that include a recognition component, as the recognizer is typically seen as an introducer of noise and researchers would like to establish both the performance of the working end-to-end system and the performance individual components would reach given perfect recognition. Performance on transcripts, then, is seen as a gold standard: the "true" accuracy of the module in question.

4.1.2 Working from recognizer output

As our goal is to incorporate nativeness classifications in a spoken language understanding system, we produced recognizer hypotheses for each read sentence and created recognizer output "documents" for each article. The word error rate (WER) for this task was 21% for the native speakers and 58% for the non-native speakers. WER is a measure which combines the results of word insertion, deletion and substitution (Lee, 1990) when the recognizer hypotheses are compared to the transcripts. An ideal speech recognizer would have a word-error rate of 0%. These system WER figures represent recognizer performance with speaker adaptation but without any specialized non-native acoustic or language modeling, as the system would not know whether the speaker was native or non-native at the time of classification.

Although any differences between the native and non-native hypotheses should be fair game for classification, it is valuable to establish whether there is something special about the way the recognizer is recognizing non-native speech, or whether the classifier is only picking up on the higher word error. Therefore, we produced a second set of hypotheses for the shared article in which WER was artificially increased by adding white noise to the speech signal. The error rate of these hypotheses was 56%.

4.2 Spontaneous speech

For spontaneous speech experiments, we used the full set of 31 Japanese, 6 Chinese, and 8 native English speakers. All speakers were executing the task described in Section 2.2, namely, asking an agent questions about sightseeing in a specified city. In this respect, the data is all compatible. However, there was some overlap in the proper nouns (names of sights, cities, restaurants, and the like) that appeared in the data. In order to ensure that the classifier was not modeling distribution of place names, we repeated all classification experiments using part-of-speech tags instead of word tokens.

With data from three native speaker groups available, we were able to evaluate 3-way classification accuracy, as well as binary classification with more than one native speaker group represented in the non-native data. Additionally, we tested binary classification accuracy of Chinese and Japanese speakers.

4.3 Feature Selection

Other machine learning and document classification tasks have been shown to benefit from feature selection. In addition, the related tasks of author and genre identification have been shown to work well using only stopwords, a form of feature selection in

Document source	word	POS
shared article (trans)	83%	74%
shared article (rec)	**94**	**100**
shared article (high-WER rec)	66	77
unique articles (trans)	41	40
unique articles (rec)	47	**77**
train=u;test=s (trans)	56	56
train=u;test=s (rec)	56	**95**
train=s;test=u (trans)	56	56
train=s;test=u (rec)	56	**83**

Table 1: Classification accuracy of read speech for two-way classification of Japanese and American English speakers reading texts in English. Baseline is 56%.

Classes	word	POS	POSNoun
native/japanese	90%	84%	**97%**
native/chinese	100	100	**100**
native/japanese/chinese	90	74	**89**
native/japanese/chinese	89	83	89 ($n \leq 3$)
native/all non-native	87	76	**96**
native/all non-native	96	90	98 ($n \leq 3$)
japanese/chinese	93	86	**100**
japanese/chinese	86	80	**100** ($n \leq 3$)

Table 2: Classification accuracy of spontaneous speech. Baselines are 83% for 2-way native/Japanese and Chinese/Japanese decisions, 72% for a 2-way native/non-native decision with Chinese speakers included in the non-native set, and 72% for a 3-way decision. Trigrams were only used where marked with ($n \leq 3$)

itself. It may be the case that non-native utterance detection will also perform well with vocabulary restricted to stopwords, or some other kind of feature selection. So while most of our experiments used no feature selection and included all stopwords, we also experimented with two forms of feature selection: pre-filtering of vocabulary by restricting it to words on one of several standard stopword lists; and feature selection by information gain score (Quinlan, 1986). We compared restricting the vocabulary to that from three stopword lists: the short SMART stopword list (48 words) the long SMART stopword list (524 words), and Mosteller and Wallace's (1984) first 70 function words. When performing feature selection by information gain, the top-ranking features were selected according to their information gain on the training set, for each train-test split.

5 Classification Results

We structured our experiments around the five contrasts discussed in Section 4:

- read versus spontaneous speech,
- transcriptions versus recognizer hypotheses,
- unique articles versus shared article,
- word tokens versus part-of-speech (POS) tokens,
- binary classification versus multi-way classification.

Tables 1 and 2 show classification results for read and spontaneous speech, representing different combinations of the remaining four conditions.

For the read speech shown in Table 1, the baseline accuracy is 56% (achieved by always classifying a document as non-native). When training and testing with the shared article, classification accuracy is always higher than the baseline. For the transcriptions, accuracy using word tokens is higher than accuracy using part-of-speech tokens, while for the recognizer hypotheses part-of-speech tokens outperform word tokens. The most surprising observation is that classification accuracy of recognizer hypotheses, both word and part-of-speech, is extremely high. Because the classification accuracy decreases when the native signal is corrupted to yield similar word error rates, we can infer that the classifier is learning something about the character of poorly recognized utterances that helps to identify non-native utterances. The fact that classification accuracy is still significantly higher than chance for the WER-matched condition, however, suggests that there is something special about the way non-native utterances are recognized, irrespective of overall recognizer performance.

Table 2 shows classification results for read speech. The baseline accuracy of this task (calculated by always picking the speaker set best represented in the training data) is 83% for 2-way native/Japanese and Chinese/Japanese decisions, 72% for a 2-way native/non-native decision with Chinese speakers included in the non-native set, and 72% for a 3-way decision.

Classification accuracy with words is very high in all cases. However, this is to a large extent due to proper nouns that occur in only one speaker set. Classification performance on part-of-speech tags is a stronger indicator of how well linguistic, and not topic, differences are being modeled; however, replacing all words with part-of-speech tags would hide differences in which verbs or verb-preposition sequences, for example, the speakers choose. We therefore ran a third experiment replacing only *nouns* (proper and not) with their parts of speech. Results from this experiment are given in the last column of Table 2, and show that this formulation of the classification task results in accuracies as high or higher than word-based classification while avoiding the risk of overtraining to the topic or task.

Feature selection	Number of Features	Accuracy
none	4800	47%
IG-524	524	69
SMART-524	524	88
IG-500	500	83
IG-200	200	74
SMART-524, IG-200	200	88
IG-70	70	70
M&W-70	70	87
IG-48	48	74
SMART-48	48	84

Table 3: Accuracy on speech recognition hypotheses is improved by feature selection based on information gain (IG). Even greater improvements are obtained by using just the vocabulary from stopword lists.

For most of the experiments described in this paper, only unigrams and bigrams were used in classification. This was because of the small size of the data set. Although the part-of-speech data contained far fewer unique word types and could support trigram modeling, we did not find that including trigrams resulted in significantly improved classification in most cases. For the spontaneous case, however, we did observe a difference in performance when the non-native document set was extended to include native speakers of Chinese. Results with trigrams are marked with ($n \leq 3$).

5.1 Feature Selection

For the feature selection experiments we used the mixed condition speech hypotheses, and found that feature selection can greatly improve classification accuracy for this most difficult condition, as shown in Table 3. Note that reducing vocabulary size by performing feature selection by information gain improves classification accuracy. However, restricting the vocabulary to stopwords from one of the SMART lists or Mosteller and Wallace's list has a much greater effect for the same vocabulary size. Different stopword lists performed equivalently, and performing further feature selection on the restricted vocabulary using information gain did not improve performance. The results shown here are for unigrams, which consistently performed better than bigrams when the vocabulary was restricted.

6 Discussion

We have found that transcriptions of spontaneous speech can be classified with high accuracy for both binary (native/non-native) and 3-way decisions. We have also found that *read* speech samples, which are all simple transformations of native-produced text, can be classified as native or non-native, and that recognizer output is classified more accurately than transcripts. To understand these results, it is helpful to look at the individual words and n-grams that contribute most to successful discrimination.

Native	Non-native
NMFS	the;the
the;NMFS	in;in
nineteen;hundreds	the
hundreds;now	in
hundreds	that
habitats;and	habitat;and
'll;grow	fishers

Table 4: Most discriminative word n-grams in transcripts of read speech, sorted by log-odds score.

Native	Non-native
noun(pl)	noun(sing)
determiner	preposition
noun(pl);preposition	preposition;preposition
adjective;noun(pl)	noun(sing);noun(sing)
gerund;particle	particle;preposition
noun(s);verb(3s)	cardinal#;cardinal#
noun(pl);modal	verb(past)

Table 5: Most discriminative part-of-speech n-grams in transcripts of read speech, sorted by log-odds score.

6.1 Transcriptions of read speech

Tables 4 and 5 show the words and parts of speech, respectively, that were important in discriminating between native and non-native transcripts of the shared article, sorted by log-odds score. The top word indicating native speech was *NMFS*, which was an acronym for the National Marine Fisheries Service. The native speakers always read this smoothly, while the non-native speakers often repeated and misread letters. The top n-gram for the non-native speakers, on the other hand, was a repetition of the determiner *the*. Non-native speakers frequently repeated words in their reading, possibly because they were unfamiliar with the next word. The term *nineteen hundreds* also played an important role in identifying native speech. This token was written in numerals in the text (*1900s*), and non-native speakers often did not know how to read it aloud. Whether a speaker read *habitats* or *habitat* (the correct word was *habitats*) was another clue to nativeness class. Reading errors involving singular-plural confusion were extremely common in the non-native speech, and relatively rare in the native speech.

The singular-plural distinction was also important in discriminating based on part of speech. A large number of plural nouns was found to be the primary indicator of nativeness. It is important to keep in mind at this point that speakers were all reading the same article, so the higher frequency of plural nouns does not necessarily indicate a preference on the part of native speakers for plural nouns, but rather a ten-

Native	Non-native
the	that
salmon	and
will	to
with	it
salmons	we
the;NMFS	someone
habitats	some

Table 6: Discriminative word n-grams in recognizer hypotheses of read speech.

Native	Non-native
noun(pl)	verb(past)
noun(pl);preposition	personal pronoun
adjective;noun(pl)	noun(sing)
noun(pl);modal	coordinating conjunction
adjective	"to"
determiner;adjective	noun(s);verb(past)
determiner;noun(pl)	personal pronoun;verb(past)

Table 7: Discriminative part-of-speech n-grams in recognizer hypotheses of read speech.

dency of non-native speakers to misread words in a text where plural nouns were frequent.

6.2 Recognizer hypotheses of read speech

Tables 6 and 7 show the important word and part-of-speech n-grams in discriminating between recognizer hypotheses of the shared read article. The most striking difference, and the one most encouraging for further work in classification of recognizer output, is the word *salmon*. This was an article about salmon populations, so this token appeared many times. In the native speech, it was generally recognized correctly. In the non-native speech, however, it was usually not, but was rather misrecognized as *some*, *someone*, and *simon*, among other words. Misrecognized native productions of the word *salmon*, on the other hand, did not tend to be misrecognized this way, but rather as the plural *salmons*, which, incidentally, is not the correct plural form and did not appear in the article but was allowed in the search because it was produced on occasion by non-native speakers.

Turning to the part-of-speech-based classification (Table 7), we can see that plural nouns continue to play a role in nativeness decisions. This is true for the noisy native data set as well as the baseline native data set. The top token on the non-native list is the past tense verb. It is not obvious why this form is so indicative of non-native speech. Past tense verbs also help to identify non-native speech in transcripts, indicating that non-native speakers are indeed on occasion reading past tense forms inappropriately, but the association is much stronger in the recognizer output. Our hypothesis is that the non-native speakers move less smoothly from word to word, and that epenthetic vowels, unnatural consonant releases, and inter-word human noise are taken by the recognizer to be a past tense ending.

Native	Non-native
am	noun(s)
proper noun	the
can;you	the;noun(s)
more	is;the
more;noun	is
give;me	noun(s);noun(s)
give	how

Table 8: Discriminative word n-grams in transcriptions of spontaneous speech.

Native	Non-native
"to";verb(base)	noun(sing)
preposition	wh-adverb
personal pronoun	verb (3s)
verb(base)	verb(3s);determiner
adjective;noun(pl)	determiner
adjective(comp.);noun(s)	wh-adverb;verb(3s)
noun(sing);modal	determiner;noun(sing)

Table 9: Discriminative part-of-speech n-grams in transcriptions of spontaneous speech.

6.3 Spontaneous speech

Discriminative tokens for spontaneous speech are given in Tables 8 and 9. The word tokens include tokens representing singular, plural, and proper nouns, avoiding overtraining on specific place names as discussed in Section 5. Because this is spontaneous speech, we are no longer looking at reading errors, but rather genuine preferences in word usage for the different speaker groups. The non-native data set consists of speakers of both Chinese and Japanese.

Nouns, specifically singular, non-proper nouns, are a strong indicator of non-nativeness. We have observed a tendency on the part of the non-native speakers to form sentences around noun phrases, saying, for example, *what is the price of the ticket of the show* where a native speaker might say *how much does the show cost*. Native speakers use more personal pronouns in their queries to the agent, as evidenced both by the importance of the personal pronoun in the part-of-speech-based classification and related verb forms like *am*. Sentences like *I'm interested in seeing the Empire State Building, can you give me more information* are common in the native data, where non-native speakers showed a strong preference for simple constructions like *how do I go to the Empire State Building*. This tendency also partly explains the importance of wh-adverbs (how, when, where, why) in identifying non-native speech.

6.4 Generalization to new data

It is clear from the results in Table 1 that the penalty for training and test condition mismatch in classification of read speech is very high. While it is possible to learn, from different renditions of a single article, a model that can classify new renditions of that same article as native or non-native, that model cannot be used, at least in this limited framework, to classify new articles – in the case of transcriptions. It appears, however, that one *can* build a model that will generalize to new data when processing recognizer output. Even when training and testing on all unique articles, a condition in which a useful model is not learned from transcriptions, when working with recognizer output a model is learned that performs significantly better than chance.

This was an unexpected and encouraging result. Because the recognizer is often viewed as a noisy channel, it was thought that some of the features that mark nativeness would be lost during speech recognition. It would appear, however, that classification of recognizer hypotheses is actually an easier task than classification of manually verified transcriptions.

7 Future Work

Based on the promising results on classification of text, we are looking at ways of incorporating this with existing techniques for identifying non-native speakers based on acoustic features, with the goal of obtaining even greater reliability from combining the two. A speech recognition system that can automatically determine whether a speaker is native or non-native, and appropriately adjust models, would ideally perform this adjustment on the fly. Thus we would also like to explore minimizing the number of words needed to be spoken before the system can identify nativeness. We are also interested in exploring a comparison between naive Bayes and language model based classification; initial experiments suggested that perplexity on a standard backoff language model may be a good discriminator for transcripts (although not as good for hypotheses, which implicitly contain some language model information). Finally, it may be helpful to apply some of the features discovered in this work to the problem of identifying native and non-native writing, which may be of relevance to writer identification.

8 Acknowledgements

We would like to thank Chris Manning, Kamal Nigam and the anonymous reviewers for helpful suggestions. The authors were supported in part by the NSF LIS Grant REC-9720374 and a Microsoft graduate research fellowship.

References

Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-based text categorization: What newspaper am I reading? In *AAAI Workshop on Learning for Text Categorization.*

S. P. Corder. 1967. The significance of learners' errors. *International Review of Applied Linguistics*, 5(4):161–170.

Pedro Domingos and Michael Pazzani. 1997. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103.

Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries, and Torsten Zeppenfeld. 1997. The Janus-RTk Switchboard/Callhome 1997 Evaluation System. In *Proc. the LVCSR Hub5-e Workshop.*

Pascale Fung and Wai Kat Liu. 1999. Fast Accent Identification and Accented Speech Recognition. In *Proc. ICASSP.*

LDC. 2000. http://www.ldc.upenn.edu.

Kai-Fu Lee. 1990. Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition. In *Proc. ICASSP.*

David Lewis. 1998. Naive (Bayes) at forty: The independence assumption. In *Proc. ECML'98.*

Laura Mayfield Tomokiyo and Susanne Burger. 1999. Eliciting Natural Speech from Non-Native users: Collecting Speech Data for LVCSR. In *Proc. the ACL-IALL Joint Workshop in Computer-Mediated Language Assessment and Evaluation in Natural Language Processing.*

Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/˜mccallum/bow.

Frederick Mosteller and David L. Wallace. 1984. *Applied Bayesian and classical inference : the case of the Federalist papers.* Springer-Verlag.

J. Ross Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proc. EMNLP.*

1987. Guide to SPEAK. Produced by the Test of English as a Foreign Language Program, Princeton, NJ.

Elaine Tarone. 1978. The phonology of interlanguage. In J.C. Richards, editor, *Understanding Second and Foreign Language Learning: Issues and Approaches.* Newbury House, Rowley, MA.

Carlos Teixeira, Isabel Trancoso, and António Serralheiro. 1996. Accent identification. In *Proc. ICSLP*, Philadelphia, PA.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *22nd Annual International SIGIR*, pages 42–49, Berkeley, CA.