DIALECT CLASSIFICATION ON PRINTED TEXT USING PERPLEXITY MEASURE AND CONDITIONAL RANDOM FIELDS

Rongqing Huang, John H.L. Hansen

Center for Robust Speech Systems
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, TX, USA
huangr@colorado.edu, john.hansen@utdallas.edu

ABSTRACT

Studies have shown that dialect variation has a significant impact in speech recognition performance, and therefore it is important to be able to perform effective dialect classification to improve speech systems. Dialects differ at the acoustic, grammar, and vocabulary levels. In this study, topic-specific printed text dialect data are collected from the ten major newspapers in Australia, United Kingdom, and United States. An n-gram language model is trained for each topic in each country/dialect. The perplexity measure is applied to classify the dialect-dependent documents. In addition to the n-gram information, further features can be extracted from text structure. Conditional Random Fields (CRF) is such a model which can extract different levels of features and is still mathematically tractable. The CRF is applied to train the language model and classify documents. Significant improvement on dialect classification is achieved by using the CRF based classifier, especially on the small size documents (10% to 22% relative error reduction). Text classification based on variable size documents is explored and a document with several hundred words is shown to be sufficient for dialect classification. The vocabulary difference among the text documents from different countries are explored and the dialect difference is smoothly connected with the vocabulary difference. Five document topics are evaluated and performance for cross topic dialect classification is explored.

Index Terms— Dialect classification, Conditional Random Fields, n-gram language model, text classification

1. INTRODUCTION

English is the native language of many countries such as Australia, Canada, United Kingdom, and United States. Each country has developed their own version of English, which is referred to as dialect or accent. A more precise definition on accent and dialect is as follows: "Accent is the cumulative auditory effect of those features of pronunciation which identify where a person is from regionally and socially. The linguistic literature emphasizes that the term refers to pronunciation only, is thus distinct from dialect, which refers to grammar and vocabulary as well" [3]. English dialects differ in the pronunciation, word selection, and grammar. Different dialects may use different words for similar meaning. For example, "lorry" vs. "truck", "lift", vs. "elevator", "rubbish" vs."trash" , "truck call" vs. "long distance call", "petrol pump" vs. "gas station", etc. have the same meaning but are used in British and American English respectively. The spelling of some words are different in British English and American English too. For example, "centre" vs. "center", "recognise" vs. "recognize", "colour" vs. "color", "behove" vs. "behoove", etc. have the same meaning (most of them have the same pronunciation too), but are used in British and American English respectively [12].

Most previous studies on dialect in the speech recognition community have focused on pronunciation differences of dialect (i.e, accent) [1, 5, 6, 8, 10, 15, 16]). In this study, we focus on the vocabulary and grammar differences of English dialects from the major English speaking countries: Australia, United Kingdom, and the United States. The vocabulary and grammar differences are seen at the text level, where n-gram language models are trained for our baseline classification algorithm. The Conditional Random Fields (CRF) algorithm [4, 9] is a relatively new technique in natural language processing. In this study, we compare the performance of n-gram language model based classifier and the CRF based classifier, and we connect the vocabulary difference with the dialect difference.

2. DATA COLLECTION

There are available corpora designed for dialect/accent analysis at the pronunciation level (e.g., [1, 16]). The speech in these corpora are either read speech with transcripts or spontaneous speech without transcripts. The read speech is based on the same read materials. Therefore, they are not suitable for vocabulary difference and grammar difference analysis. To the best of our knowledge, there are no such corpora available yet in the speech community. In order to study the word selection and grammar difference among English dialects, we collected news data drawn from the ten major newspapers from Australia (AU), Untied Kingdom (UK), and United States (US). We selected news articles from the online version of newspapers. Table 1 shows the newspapers from where we extract documents. We selected five topics which were of primary interest in the world at a certain time period. The topics include politics, military, environment,sports, and economy. The time periods when these articles were written are strictly confined, which is usually from one month to one year. The length of the articles are limited as well. If the article is shorter than 100 words, we believe they are headline news and they are excluded. We focused on collecting the articles which we felt would have vocabulary or grammar differences. We also maintained the same amount of documents across the different countries. Table 2 shows the text document information after the HTML tags are removed. There are over 8 thousand documents, 5.7

This work was supported by U.S. Air Force Research Laboratory, Rome NY under contract No. FA8750-04-1-0058. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Air Force.

million words in our collection. The average length of the documents are 700 words. The documents of each topic are randomly partitioned into equal size training and test corpora.

Table 1. *The data source: ten newspapers from AU, UK, and US*

Country/ Dialect	Newspapers
AU	Australian; Canberra Times; Courier Mail; Melbourne Herald Sun;West Australian; Sydney Daily Telegraph; Cairns Post; Mercury; Advertiser; AAP Newsfeed
UK	Daily Mail; Guardian; Mirror; Sun; London Times; London Daily Telegraph; London Evening Standard; Independent; Daily Star; News of the World;
US	Los Angeles Times; New York Times; USA Today; Washington Post; Boston Globe; Chicago Sun-Times; St. Louis Post-Dispatch; Christian Science Monitor; Miami Herald; San Diego Union-Tribune

Table 2. *Information of the corpora*

Topic	Keywords	Time	# Words	# Doc.
Climate	Global Warming Climate	11/2003 - 11/2004	1283K	2026
Election	US Election Bush Kerry	9/2004 - 11/2004	1208K	1621
Iraq	Iraq Saddam	9/2004 - 11/2004	544K	719
Oil	Petroleum Trade	11/2003 - 11/2004	857K	1379
Olympic	Olympic Athens Medal	8/2004	1727K	2587

3. DOCUMENT CLASSIFICATION ALGORITHMS

The text documents represent sequential data. The sequential classification does not fit well for classical supervised learning, which assumes that the training data is drawn independently and identically (i.e., iid) from a joint distribution $P(\mathbf{X}, y)$. In sequential classification, training data is actually a sequence of $(\mathbf{X}, y)$ pairs, and there is strong correlation across the sequence. If this correlation is ignored, the classification performance will be very poor. A well-known sequential classification algorithm is based on the n-gram language model which we consider here.

3.1. Baseline: n-gram language model based classifier

We assume that the text document is composed of many sentences. Each sentence can be regarded as a sequence of words $\mathbf{W}$. The probability of generating $\mathbf{W}$ can be measured as

$$\begin{aligned} P(\mathbf{W}|D) &= P(w_1, w_2, \ldots, w_m|D) \\ &= P(w_1|D)P(w_2|w_1, D)\cdots \\ &\quad P(w_m|w_1, w_2, \ldots, w_{m-1}, D) \\ &= \Pi_{i=1}^{m} P(w_i|w_1, w_2, \ldots, w_{i-1}, D) \\ &= \Pi_{i=1}^{m} P(w_i|w_{i-n+1}, \ldots, w_{i-1}, D), \end{aligned} \quad (1)$$

where m is the number of words in $\mathbf{W}$, w_i is the word, and $D \in \{\text{AU, UK, US}\}$ is the dialect specific language model. The final equation comes from the n-gram definition. The n-gram probabilities are calculated from occurrence counting. The final classification decision is,

$$C = \arg\max_{D} \Pi_{\mathbf{W}\in\Omega} P(\mathbf{W}|D), \quad (2)$$

where Ω is the set of sentences in a document and $D \in \{\text{AU, UK, US}\}$. The prior probabilities are omitted since these classes are assumed having equal priors.

In this study, we use the derivative measure of the cross entropy known as the test set perplexity for dialect classification. If the word sequence is sufficiently long, the cross entropy of the word sequence $\mathbf{W}$ is approximated as

$$H(\mathbf{W}|D) = -\frac{1}{m}\log_2 P(\mathbf{W}|D), \quad (3)$$

where m is the length of the test word sequence $\mathbf{W}$ measured in words. The perplexity of the test word sequence $\mathbf{W}$ as it relates to the language model D is,

$$PP(\mathbf{W}|D) = 2^{H(\mathbf{W}|D)} = (P(\mathbf{W}|D))^{-\frac{1}{m}}. \quad (4)$$

The perplexity of the test word sequence is the generalization capability of the language model. The smaller the perplexity, the better the language model generalizes to the test word sequence. The final classification decision is,

$$C = \arg\max_{D} \Pi_{\mathbf{W}\in\Omega} (P(\mathbf{W}|D))^{-\frac{1}{m}}, \quad (5)$$

where Ω is the set of sentences in a document, $D \in \{\text{AU, UK, US}\}$, and m is the length of a sentence. Comparing Eq. 2 with Eq. 5, we find that the perplexity measure is actually the normalized probability measure and the normalization factor is the sentence length.

3.2. Conditional Random Fields (CRF) based classifier

The n-gram model is easy to implement and can achieve good performance for language modeling and dialect classification. However, there are further information in the text than n-gram probabilities. The conditional model is such a model which can learn more features than n-gram models. For sequential data, let $\mathbf{X}$ be the random variable of the input features, $\mathbf{Y}$ be the random variable of the sequence of labels. The probabilistic model can be used to represent the relationship between $\mathbf{X}$ and $\mathbf{Y}$. The conditional models attempt to learn $P(\mathbf{Y}|\mathbf{X})$ instead of the joint probability $P(\mathbf{X}, \mathbf{Y})$, which is learned in a generative model such as the n-gram model and Hidden Markov Model (HMM). The Conditional Random Fields (CRF) [9, 13] are the popular conditional models. Table 3 lists the comparison between the typical generative model (i.e., HMM) and the typical conditional model (i.e., CRF).

The CRF model for each class is defined as [9, 13]

$$P(\mathbf{Y}|\mathbf{X}) \equiv P_{\boldsymbol{\lambda}}(\mathbf{Y}|\mathbf{X}) = \frac{\exp(\boldsymbol{\lambda} F(\mathbf{Y}, \mathbf{X}))}{\sum_{\mathbf{y}} \exp(\boldsymbol{\lambda} F(\mathbf{y}, \mathbf{x}))}, \quad (6)$$

$$\text{where } F(\mathbf{y}, \mathbf{x}) = \sum_i f(\mathbf{y}, \mathbf{x}, i).$$

Here, $f(\mathbf{y}, \mathbf{x}, i)$ is the feature at position i and $\boldsymbol{\lambda}$ is the parameter vector of the CRF. Any real value function can be used for $f(\mathbf{y}, \mathbf{x}, i)$. but in practice, boolean functions are used for simplicity. Sample boolean features are "$\mathbf{x}$ begin with a number: 1(yes), 0(no)", "$\mathbf{x}$ ends with a exclamation mark: 1(yes), 0(no)", etc. The n-gram information can be easily added through the feature functions as well.

Table 3. *HMM vs. CRF*

Model	Learn	Feature (X)	Strengths	Weaknesses
HMM	$P(\mathbf{X}, \mathbf{Y})$	Local, and independent only	Elegant interpretation: explains how $\mathbf{X}$ is generated	Strong assumptions: features are localized and independent
CRF	$P(\mathbf{Y}\|\mathbf{X})$	Arbitrary	Encodes any level of information, discriminative and global training	Computational expensive, especially in training

The size of the parameter vector will equal to the number of features in the model. The likelihood of a given sample $(\mathbf{x}, \mathbf{y})$ is,

$$L = \log P_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) \tag{7}$$

The parameter estimation of CRF is based on Maximum Likelihood (ML) principle. Assume there are K training samples, with the total likelihood of the training data given by,

$$L_{\boldsymbol{\lambda}} = \sum_{k=1}^{K} \log P_{\boldsymbol{\lambda}}(\mathbf{y}_k|\mathbf{x}_k). \tag{8}$$

So the ML estimation of the parameter vector $\boldsymbol{\lambda}$ of the CRF is,

$$\bigtriangledown L_{\boldsymbol{\lambda}} = \sum_{k=1}^{K} \Big\{ F(\mathbf{y}_k, \mathbf{x}_k) - \mathrm{E}_{P_{\lambda}(\mathbf{Y}|\mathbf{x}_k)}[F(\mathbf{Y}, \mathbf{x}_k)] \Big\} \equiv 0, \tag{9}$$

where E is the expectation operation. From Eq. 9, we observe that the ML estimation of the CRF is the same as the Maximum Entropy (ME) estimation of the model, which assures that the model is unbiased [2]. This is an attractive mathematical property, since the ML estimation is usually not the same as ME estimation. Eq. 9 can be solved via the Limited-Memory Quasi-Newton (L-BFGS) algorithm [11]. During evaluation, Eq. 7 is computed as the likelihood score of a given test sample to the CRF model, and the final decision is,

$$C = \arg\min_{D} \log P_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}, D), \tag{10}$$

where $D \in \{\text{AU, UK, US}\}$.

4. EXPERIMENTS

In this section, we first explore dialect classification using the n-gram language model based classifier and the CRF based classifier, then cross topic classification and the vocabulary differences in the English countries.

4.1. Document classification

For the baseline system, we build the n-gram language model. The classification is based on the perplexity of documents as reflected in Eq. 5, where the OOV (Out of Vocabulary) is considered in the perplexity computation.

The first experiment is to determine how many grams are needed for dialect classification. Table 4 shows the classification performance of n-gram language models in the topics as "n" increases. The Katz backoff smoothing [7] is applied in model training, and the discounting strategy for n-gram model training is Witten Bell discounting [14]. The cutoffs are set to one (i.e., only n-grams occur more than one time are counted). From Table 4, we observe that 2-gram language models achieve best performance overall, so the following experiments are based on 2-gram models for the n-gram based classifiers.

Table 4. *Classification accuracy (%) of the n-gram language models*

Topic/ # of Gram	1	2	3	4	5
Climate	41.7	88.3	87.6	87.8	88.0
Election	55.5	83.2	81.8	81.9	81.8
Iraq	54.4	82.1	80.7	81.0	80.7
Oil	46.0	91.0	91.0	90.5	91.0
Olympic	61.3	91.8	91.4	91.3	91.1

Table 5 shows the dialect classification accuracy for the proposed CRF based classifier on variable size documents in the five topics versus the n-gram based classifier. The number outside the parentheses is the classification accuracy of the n-gram based classifier; the number inside the parentheses is the classification accuracy of the CRF based classifier. The average number of words for the original documents are 700 for all topics. Fig. 1 shows the variable size document classification accuracy averaged on the five topics using the n-gram based classifier and the CRF based classifier. Table 5 and Fig. 1 clearly show that the CRF based classifier significantly outperforms the n-gram based classifier especially on small size documents. An obvious explanation is that the CRF encodes many levels of features including the n-gram information. It is not a surprise that it can outperform the model which uses only n-gram information. The other observation is that the 300-word document is sufficient for effective dialect classification.

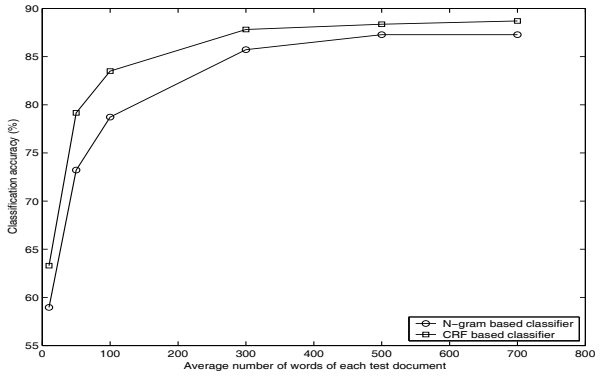

Fig. 1. *Dialect classification accuracy averaged on the five topics using the n-gram based classifier and the CRF based classifier.*

4.2. Cross topic evaluation

It is interesting to evaluate how related the topics are and how much degradation there is if the topic is mismatched during dialect classification. Table 6 shows the performance for a cross topic evaluation. The classifier used here is the n-gram based classifier. The test document is the original document. From Table 6, we observe that classification performance degrades severely when the language models are trained using different topic data. However, the perfor-

Table 5. *Dialect classification accuracy (%) on variable size documents using the n-gram based classifier and the CRF based classifier. The numbers outside and inside the parentheses are the classification accuracy of the n-gram based classifier and the CRF based classifier respectively.*

Topic/Size	10 Words	50 Words	100 Words	300 Words	500 Words	Original Document (700 Words)
Climate	57.6 (61.9)	71.2 (76.6)	77.7 (81.3)	85.7 (86.3)	87.7 (87.4)	88.3 (86.9)
Election	53.5 (56.5)	66.9 (72.6)	72.9 (78.2)	81.6 (83.9)	83.0 (85.2)	83.2 (85.9)
Iraq	54.8 (60.0)	67.4 (74.4)	72.5 (78.3)	79.3 (82.8)	81.7 (82.3)	82.1 (82.5)
Oil	62.5 (67.0)	78.6 (84.2)	83.5 (88.9)	89.6 (93.2)	91.1 (93.8)	91.0 (94.9)
Olympic	66.4 (71.1)	82.0 (88.0)	87.0 (90.8)	92.4 (92.9)	92.9 (93.1)	91.8 (93.3)

mance is almost the same as using the topic-specific language models when using the "Election" language models to evaluate the "Iraq" data (79.6% vs. 82.1%). The "Iraq" was one of the most important issues in the US presidential election, so this is reasonable.

Table 6. *Cross topic dialect classification evaluation*

Test/Train	Climate	Election	Iraq	Oil	Olympic
Climate	88.3	58.5	65.8	58.6	53.2
Election	37.6	83.2	63.3	36.4	52.7
Iraq	57.5	**79.6**	**82.1**	43.3	57.5
Oil	81.3	61.8	75.0	91.0	64.8
Olympic	67.8	63.2	56.3	51.5	91.8

4.3. Dialect distance based on vocabulary difference

Next, we consider the vocabulary differences among the dialects. The 500 most frequently used words, excluding function words such as "the", "a", "is", "he" and so on, are considered. The inflectional variants such as "-ing", "-s" and "-ed" are considered and the stem of the words are merged. Table 7 shows the percentage of words which are different between the dialects.

Table 7. *Vocabulary difference (%) of the 500 most frequent words*

Topic	AU–UK	UK–US	AU–US
Climate	22.8	25.8	29.0
Election	20.0	20.8	25.2
Iraq	32.2	32.4	33.8
Oil	27.4	24.0	34.8
Olympic	30.8	35.0	36.0
Average	26.6	27.6	31.8

From Table 7, the larger the vocabulary difference, the larger of the distance between the two dialects. Therefore, the distance between Australian and American English is the largest, followed by British and American English. The most similar dialects, having the smallest distance between them, is Australian and British English. This observation is consistent with our perceived linguistic knowledge. Another interesting observation is that if the topic is country-specific, such as the the US presidential election, the vocabulary difference is small; if the topic is global, such as the 2004 Olympic games, the vocabulary difference is large. We believe Australian people tend to "copy" what Americans say in their presidential election in the newspaper, and people tend to develop their own interests or focus in the Olympic games.

5. CONCLUSIONS

In this study, we showed that topic-specific printed text documents from the English countries can be classified across dialects. The dialect classification accuracy is from 82% to 95% in our study. The CRF based classifier significantly outperforms the n-gram based classifier in dialect classification especially on small size documents (10% - 22% relative error reduction). If the language models are topic mismatched, the classification performance is degraded severely. The vocabulary difference is also a good indicator of dialect difference in English countries. The vocabulary difference is large between Australian and American English, and is small between Australian and British English. This observation is consistent with our perceived linguistic knowledge. It is suggested that the text based knowledge gained from this study would provide insight into a combined acoustic/text based dialect classification system.

6. REFERENCES

[1] P. Angkititrakul, and J.H.L. Hansen, "Advances in phone-based modeling for automatic accent classification", *IEEE Trans. Speech and Audio Processing*, vol. 14, pp. 634-646, 2006

[2] A. L. Berger, S. A. Della Pietra and V. J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing", in *Computational Linguistics*, vol. 22, no. 1, pp. 39-71, 1996

[3] D. Crystal, "A Dictionary of Linguistics and Phonetics", Blackwell Publishers, Malden, MA. USA, 1997

[4] T. G. Dietterich, A. Ashenfelter, Y. Bulatov, "Training Conditional Random Fields via Gradient Tree Boosting", *ICML*, 2004

[5] R. Huang and J.H.L. Hansen, "Dialect/Accent Classification via Boosted Word Modeling", *ICASSP-05*, Philadelphia, 2005

[6] R. Huang and J.H.L. Hansen, "Advances in Word based Dialect/Accent Classification", *Proc.InterSpeech-05*, Lisbon, Portugal, 2005

[7] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer", *IEEE Trans Acoustics, Speech, and Audio Processing*, vol. 3, pp. 400-1, 1985

[8] K. Kumpf and R. W. King, "Foreign Speaker Accent Classification using Phoneme-Dependent Accent Discrimination Models and Comparisons with Human Perception Benchmarks", *Eurospeech-97*, 1997

[9] J. Lafferty, A. McCallum, F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", *ICML-01*

[10] D. Miller and J. Trischitta, "Statistical Dialect Classification Based on Mean Phonetic Features", *ICSLP*, 1996

[11] J. Nocedal and S. J. Wright, "Numerical Optimization", *Springer*, 1999

[12] P. Peters, "The Cambridge Guide to English Usage", Cambridge University Press, UK, 2004

[13] F. Sha and F. Pereira, "Shallow Parsing with Conditional Random Fields", in *Proc. Human Language Technology-NAACL*, Edmonton, Canada, 2003

[14] I. H. Witten and T. C. Bell, "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression", *IEEE Trans. Information Theory*, 37(4), pp. 1085-94, 1991

[15] Q. Yan and S. Vaseghi, "Analysis, Modeling and Synthesis of Formants of British, American and Australian Accents", *ICASSP*, 2003

[16] M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", in *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 31-44, 1996