Automatic Estimation of Dialect Mixing Ratio for Dialect Speech Recognition

Naoki Hirayama[1], *Koichiro Yoshino*[1], *Katsutoshi Itoyama*[1], *Shinsuke Mori*[1,2], *Hiroshi G. Okuno*[1]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]Academic Center for Computing and Media Studies, Kyoto University, Japan
hirayama@kuis.kyoto-u.ac.jp, yoshino@ar.media.kyoto-u.ac.jp, itoyama@kuis.kyoto-u.ac.jp,
forest@i.kyoto-u.ac.jp, okuno@i.kyoto-u.ac.jp

Abstract

This paper proposes methods for determining an appropriate mixing ratio of dialects in automatic speech recognition (ASR) for dialects. To handle ASR for various dialects, it has been reported to be effective to train a language model using a dialect-mixed corpus. One reason behind this is geographical continuity of spoken dialect; we regard spoken dialect as a mixture of various dialects. This mixing ratio changes at every moment as well as depends on a speaker. We can improve recognition accuracy by giving an appropriate dialect mixing ratio for a speaker's dialect. The mixing ratio is generally unknown and requires to be estimated and updated referring to input utterances. We handle two methods for updating it based on recognition results; one is to compute contribution of dialects for each recognized word, and the other is to predict mixture information referring to a whole recognized sentence based on topic modeling. The experimental result shows that the mixing ratio estimated by these methods realized higher recognition accuracy than a fixed mixing ratio.

Index Terms: dialect, supervised latent Dirichlet allocation (sLDA), mixing ratio.

1. Introduction

Speech recognition and dialogue systems have been recently embedded in various devices, such as smartphones. Since these devices might be used by many people, speech recognition systems should recognize the utterances of as many speakers as possible. Voices have different characteristics, such as age, speech rate, accent, and vocabulary. Nevertheless, most of these systems do not handle various characteristics; they usually assume adult speakers, speech rate of reading, and vocabulary in written language. Recognition accuracy will drastically deteriorate for spontaneous speech [1] that has different characteristics from the assumed ones. In this paper, automatic speech recognition (ASR) of dialects in particular is handled.

We handle automatic estimation of dialect mixing to improve recognition accuracy. We regard a spoken dialect as a mixture of various dialects, the mixing ratio of which changes at every moment as well as depending on the speaker. In our method based on dialect mixing [2], first, dialect-specific pronunciation dictionaries are created, and then the probabilities given to each pronunciation are weightedly averaged. Features of dialects can be divided into three types: (1) pronunciation [3, 4], (2) vocabulary [5], and (3) word order [6]. One famous example of the first type is the set *"marry, merry, and Mary"*. One example of the second type is *"mind the gap"* and *"watch your step"*, both of which are a warning for boarding trains. The third type has the example of *"next Tuesday"* and *"Tuesday next"* in Canadian dialect. Our method covers the first and second types and ignores the third type to simplify the problem. One more restriction is that the phone sets of dialects are all equal, which enables the first and second types to be treated equally as difference in phoneme sequences. Our method requires an appropriate dialect mixing ratio for a speaker's dialect to improve recognition accuracy, mainly because a spoken dialect is geographically continuous [7, p.71] due to the movement of people between areas. Various dialects with small, differing characteristics are spoken even in a small area. Dialects of Figure 1 are one of the examples, though they are all roughly categorized into Kansai dialect. In this paper, Japanese standard language is termed common language (CL).

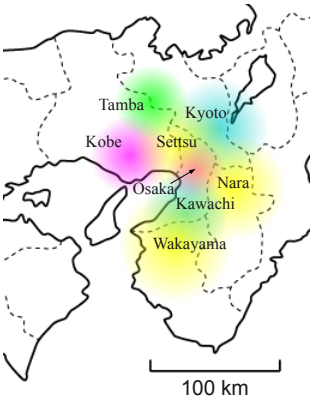


Figure 1: Many dialects concentrated in Japan. Each dialect is named after the city or area where it is spoken.

We should determine which words are specific to a dialect or widely used regardless of dialects. Our methods are twofold. The first one is to simply count dialect-specific words and calculate the dialect mixing ratio. Dialect-specific words can be determined by referring to the pronunciation dictionaries in [2]. If some pronunciation appears in only one dictionary for a dialect, it is regarded as specific to the dialect. The second one is to model dialects and their vocabulary that appears in a sentence. One of the most general models is a topic model, where each topic has its own distribution of words. We adopt the supervised latent Dirichlet allocation (sLDA) [8], a kind of supervised topic models, to categorize words into topics with different dependencies on dialects. We regard the response variable of sLDA as a real value that represents the dialect mixing ratio.

This paper is organized as follows. Section 2 reviews related work on dialect ASR. Section 3 summarizes the targeted

ASR system. Section 4 discusses our methods to determine the dialect mixing ratio. Section 5 describes our evaluation of the system. Section 6 concludes this paper and states future work.

2. Related Work

Most previous studies handling dialect ASR require huge amounts of data that cannot be practically collected or do not treat dialects systematically.

Ching et al. [9] described the acoustic properties of the Cantonese dialect of the Chinese language, such as on the basis of energy profiles, pitch, and duration. Miller et al. [10] studied the discrimination of Northern and Southern US dialects on the basis of the phonetic features. These studies require large amounts of speech data to train the distribution of features as well as taking no account of differences in vocabulary [11].

Lyu et al. [12] developed an ASR system for Chinese dialects that uses a hand-written character-to-pronunciation mapping. This has two disadvantages: the cost of developing the mapping and the difficulty in extending it to dialect mixing.

To develop [2] dialect ASR systematically trained on a realistic data set, we introduced statistical methods of emulating a linguistic corpus for training a language model by using a large linguistic corpus of CL (common language) and a small CL-dialect parallel corpus. The system was able to recognize utterances in multiple dialects or their mixture, but it had to be given an appropriate ratio of a dialect mixture for a speaker. This paper deals with estimating the ratio of a dialect mixture.

3. Dialect ASR and Mixing Dialects

We summarize the dialect speech recognition system described in [2]. This system targets Japanese dialects, but its structure itself does not depend on languages, though it assumes that the word order does not change.

Practically large linguistic corpora in dialects are not available. We emulate a large dialect corpus to build a statistically reliable language model by transforming a large CL corpus. The transformation is conducted by using a phoneme-sequence transducer from the CL to a dialect, which was developed by using a CL-dialect parallel corpus and converts a pronunciation in the CL to that in the dialect. The phoneme-sequence transducer is modeled as a weighted finite-state transducer (WFST) [13], which outputs multiple candidates together with their probability. Namely, it handles word boundaries as well as pronunciation in an input sequence, and outputs word-wise pronunciation in a specific dialect. We create a pronunciation dictionary in an ASR system that determines the pronunciation of each word, referring to the output. Let $\#(x)$ be the number of CL word x that appears in the original sentences and $\#(y|x)$ be the number of pronunciations y given to word x. Then, the pronunciation probability given a word, namely in-class probability where each class corresponds to a word, $P_c(y|x)$, is written as

$$P_c(y|x) = \frac{\#(y|x)}{\#(x)} = \frac{\#(y|x)}{\sum_y \#(y|x)}. \quad (1)$$

This method can handle only one dialect alone. The following is the way of handling multiple dialects. We compute the weighted (arithmetic) mean of in-class probability over targeted dialects. Let P_c of Equation (1) for dialect d be rewritten as $P_{c,d}$, then

$$P_{c,mix}(y|x) = \sum_d r_d P_{c,d}(y|x), \quad (2)$$

$$\text{s.t.} \quad \sum_d r_d = 1,\ r_d \geq 0$$

gives the in-class probability for a dialect mixture, which contains all kinds of pronunciations that appear in a pronunciation dictionary for any dialect.

In the following sections, we assume that $P_{c,d}(y|x)$ is already computed and discuss how to determine the weights r_d.

4. Dialect Ratio Estimation

In this section, we describe the model for the dialect mixture and how to estimate the ratio on the basis of the model. The following are two kinds of estimation methods.

4.1. Simple Counting

This method involves pronunciation dictionaries to estimate the dialect mixing ratio of each sentence. Let D be the number of mixed dialects and N be the length (the number of words) of a recognized sentence. A recognized sentence can be represented as a pair of $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$, where x_n represents a word entry and y_n is one of its pronunciations. Since we have no knowledge about a speaker's dialect beforehand, we begin from the equal weights of $r_d = 1/D$. For each recognized word x_n and y_n, the dialect mixing ratio is calculated as

$$r'_{d,n} = \frac{r_d P_{c,d}(y_n|x_n)}{\sum_{d'} r_{d'} P_{c,d'}(y_n|x_n)}, \quad (3)$$

where $P_{c,d}$ is defined as P_c in Equation (1). The value of $r'_{d,n}$ ranges from zero to one; it will be zero if the pronunciation y_n (given x_n) does not appear in the pronunciation dictionary of dialect d and one if y_n appears only there. Given all $r'_{d,n}$, we estimate the dialect weights of a sentence as

$$r'_d = \frac{1}{N} \sum_{n=1}^{N} r'_{d,n}. \quad (4)$$

Equations (3) and (4) play a role in updating dialect weights r_d to r'_d. If more dialect-specific pronunciations appear in $\boldsymbol{y}$, the value of r'_d will be much larger.

4.2. Dialect Modeling based on Topic Model

The second method of the estimation is dialect modeling based on topic model. Some words appear in a sentence regardless of a speaker's dialect, while other words appear in a specific dialect only. In other words, a speaker's dialect determines the distributions of word frequencies.

We conduct modeling on the basis of supervised latent Dirichlet allocation (sLDA) [8], which is an extended version of latent Dirichlet allocation (LDA) [14], a kind of topic model. In LDA and sLDA, each topic has its own distribution of word appearances, and each word in a sentence is sampled from one of the topics. We regard the response variable of sLDA as the dialect mixing ratio. The topic of each word is also sampled from a topic distribution stemming from Dirichlet distribution. In sLDA, pairs of a sentence and its score are modeled and the score of a new sentence is determined by using the estimated topics of each word in the sentence. In the application in this paper, each topic corresponds to a group of words that have a similar dependency on dialects. We regard the pairs of a word and its pronunciation in a recognized sentence as a document in the context of sLDA.

4.2.1. Formalization of sLDA

Let D and W be the number of given documents and the size of the vocabulary, respectively. Additionally let N_d be the length (the number of words) of document d ($d = 1, 2, \ldots, D$). Each word w_d of document d is a pair of x and y in Section 3. The number of topics K is assumed to be given here. The generative process for a document is as follows, where the regression error of a response variable follows a Gaussian distribution. For each document d,

1. Draw topic proportions $\boldsymbol{\theta}_d|\boldsymbol{\alpha} \sim \mathcal{D}(\boldsymbol{\alpha})$.
2. For each word w_{dn}, the n-th word of document d,
   (a) Draw topic assignment $z_{dn}|\boldsymbol{\theta}_d \sim \mathcal{M}(\boldsymbol{\theta}_d)$.
   (b) Draw word $w_{dn}|z_{dn}, \boldsymbol{B} \sim \mathcal{M}(\boldsymbol{B}_{z_{dn}})$.
3. Draw response variable $y_d|\boldsymbol{z}_d, \boldsymbol{\eta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{\eta}^\mathsf{T}\bar{\boldsymbol{z}}_d, \sigma^2)$, where
$$(\bar{\boldsymbol{z}}_d)_k = \bar{z}_{dk} = \frac{1}{N_d}\sum_{n=1}^{N_d}\delta_{z_{dn},k}, \tag{5}$$
   and $\delta_{i,j}$ is the Kronecker delta.

$\mathcal{D}, \mathcal{M}$, and $\mathcal{N}$ denote a Dirichlet distribution, a multinomial distribution, and a normal distribution, respectively. Hyperparameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ determines the likelihood of a topic distribution. Hyperparameter $\boldsymbol{B}$ is a $K \times W$ matrix, where $\boldsymbol{B}_k = (\beta_{k1}, \ldots, \beta_{kW})$ is a word distribution of topic k. Topic assignment z_{dn} ($n = 1, 2, \ldots, N_d$) is the index of the topic assigned to w_{dn}, the n-th word of document d. Parameter $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ determines the influence of topics on the response, and σ^2 determines the variance of regression errors.

4.2.2. Training

We adopt a variational Bayesian method to approximate the topic distribution $\boldsymbol{\theta}_d$ by that with respect to $\boldsymbol{\gamma}_d$ and the topic assignment z_{dn} by that with respect to $\boldsymbol{\phi}_{dn}$. The training process comprises two steps. First we estimate variational parameters $\boldsymbol{\gamma}_d$ and $\boldsymbol{\phi}_{dn}$, and then we update distribution parameters $\boldsymbol{\eta}$ and σ^2 and hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{B}$. These two steps are repeated iteratively until the parameters converge. All elements of $\boldsymbol{\alpha}$ are fixed to $50/K$ to simplify the training.

We excerpt only parameter update laws in the following; see [8, 14] for their derivation. Update laws are represented by using $\boldsymbol{\gamma}_d$ and $\boldsymbol{\phi}_{dn}$ instead of $\boldsymbol{\theta}_d$ and $\boldsymbol{z}_d$.

E-step Parameters ϕ_{dni} and γ_{di} are updated until they converge. ϕ_{dni} is the probability that the n-th word in document d belongs to topic i, which is normalized so that $\sum_{k=1}^{K}\phi_{dnk} = 1$. The E-step is conducted document-wise.

$$\phi_{dni} \propto \beta_{iw_{dn}} \exp\left(\Psi(\gamma_{di}) + \frac{y_d}{N_d\sigma^2}\eta_i - \frac{1}{2N_d^2\sigma^2}\left(2(\boldsymbol{\eta}^\mathsf{T}\boldsymbol{\phi}_{d,-n})\eta_i + \eta_i^2\right)\right), \tag{6}$$

$$\gamma_{di} \leftarrow \alpha_i + \sum_{n=1}^{N_d}\phi_{dni}, \tag{7}$$

where Ψ denotes the digamma function, namely $\Psi(x) = \frac{\partial}{\partial x}(\ln\Gamma(x)) = \Gamma'(x)/\Gamma(x)$, and $\boldsymbol{\phi}_{d\bar{n}} = \sum_{m\neq n}\boldsymbol{\phi}_{dm}$.

M-step Parameters $\boldsymbol{\eta}, \sigma^2, \boldsymbol{\alpha}$ and $\boldsymbol{B}$ are updated. Each update is conducted only once for each M-step. Parameter β_{ij} is normalized so that $\sum_{j=1}^{W}\beta_{ij} = 1$.

$$\boldsymbol{\eta} \leftarrow \left(\sum_{d=1}^{D}\frac{1}{N_d^2}\sum_{n=1}^{N_d}\left(\boldsymbol{\phi}_{dn}\boldsymbol{\phi}_{d\bar{n}}^\mathsf{T} + \mathrm{diag}\{\boldsymbol{\phi}_{dn}\}\right)\right)^{-1}\boldsymbol{E}\boldsymbol{y}, \tag{8}$$

$$\sigma^2 \leftarrow \frac{1}{D}\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - \boldsymbol{y}^\mathsf{T}\boldsymbol{E}\boldsymbol{\eta}\right), \tag{9}$$

$$\beta_{ij} \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}\delta_{w_{dn},j}, \tag{10}$$

where

$$\boldsymbol{E} = \begin{bmatrix}\frac{1}{N_1}\sum_{n=1}^{N_1}\boldsymbol{\phi}_{1n} & \cdots & \frac{1}{N_D}\sum_{n=1}^{N_D}\boldsymbol{\phi}_{Dn}\end{bmatrix}. \tag{11}$$

4.2.3. Prediction

Once the training process is finished, we predict the response value to an input sentence $\boldsymbol{w} = (w_1, w_2, ..., w_N)$. We update parameters $\hat{\phi}_{ni}$ and $\hat{\gamma}_i$ iteratively with:

$$\hat{\phi}_{ni} \propto \beta_{iw_n}\exp\Psi(\hat{\gamma}_i), \tag{12}$$

$$\hat{\gamma}_i \leftarrow \alpha_i + \sum_{n=1}^{N}\hat{\phi}_{ni}, \tag{13}$$

and predict the response value $\hat{y}$ with:

$$\hat{y} = \boldsymbol{\eta}^\mathsf{T}\left(\frac{1}{N}\sum_{n=1}^{N}\hat{\boldsymbol{\phi}}_n\right). \tag{14}$$

5. Evaluation

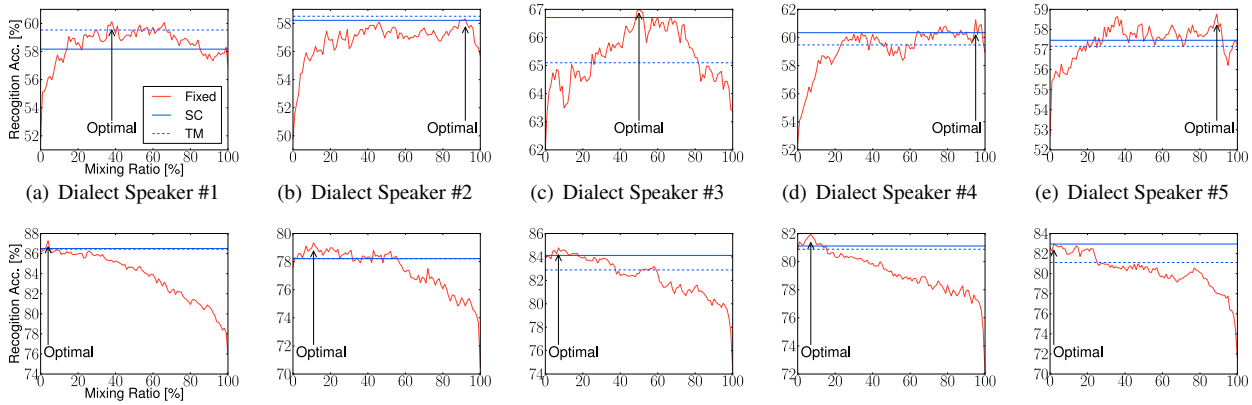
We carried out an experiment to evaluate the two methods above on speech recognition accuracy.

5.1. Conditions

We describe the training data for the phoneme-sequence transducers, language models, and acoustic models. The phoneme-sequence transducers in this experiment adopted the parallel corpus [15] of the Kansai area (Osaka, Kyoto and Hyogo Prefectures), composed of 24,597 words. Language models were trained on sentences of 3,000,000 questions and corresponding answers (71.2 million words) in the Yahoo! Q&A corpus (daily-life category). To exclude noise such as Internet slang, the sentences were chosen with entropy-based filtering [16] by using the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [17]. The vocabulary of language models comprised words that appeared more than ten times in the sentences, and the size of vocabulary was 42,845. Acoustic models were trained on 70.2 hours of talking altogether by 500 speakers in the Corpus of Spontaneous Japanese (CSJ) [18] and 23.3 hours of talking altogether by 308 speakers in Japanese Newspaper Article Sentences (JNAS) [19].

For evaluation, five Kansai dialect and five CL speakers read 100 common sentences, where Kansai dialect speakers translated the sentences into their natural dialect before reading. We adopted Julius [20] as the ASR engine in this experiment.

We adopted 1/100 of the training data for a language model of CL and the Kansai dialect as the training corpus for sLDA. CL sentences were given a response value of 0.0, and the Kansai dialect sentences were given 1.0. The prediction of the response value for a recognized sentence was the mixing ratio used in



(a) Dialect Speaker #1 (b) Dialect Speaker #2 (c) Dialect Speaker #3 (d) Dialect Speaker #4 (e) Dialect Speaker #5

(f) CL Speaker #1 (g) CL Speaker #2 (h) CL Speaker #3 (i) CL Speaker #4 (j) CL Speaker #5

Figure 2: Word recognition accuracy of five subjects for changing fixed mixing ratios (Fixed: red curve) versus two mixing ratios automatically controlled by simple counting (SC: solid blue line) and by topic modeling (TM: dashed blue line). The horizontal axis denotes the mixing ratio [%] of a dialect, and the vertical axis denotes word recognition accuracy [%].

Table 1: Word recognition accuracy [%] for ratios automatically controlled, the optimal fixed mixing ratio, and the fixed mixing ratio of 100% (dialect) or 0% (CL). SC and TM stand for simple counting and topic modeling, respectively.

(a) Dialect speakers

	#1	#2	#3	#4	#5
SC	58.2	58.2	**66.7**	**60.3**	**57.5**
TM	**59.5**	**58.5**	65.1	59.5	57.2
Optimal	60.1	58.3	67.0	61.3	58.8
Ratio = 100%	57.1	55.7	63.4	59.0	57.2

(b) CL speakers

	#1	#2	#3	#4	#5
SC	**86.5**	78.2	**84.1**	**81.1**	**83.0**
TM	86.4	78.2	82.9	80.9	81.1
Optimal	87.3	79.3	84.8	81.9	83.0
Ratio = 0%	86.2	77.6	84.1	80.9	82.6

recognition of the next utterance. If the response value was less than 0.01 or more than 0.99, we regarded it as 0.01 and 0.99, respectively. The number of topics K was fixed to 12.

5.2. Results

Figure 2 shows the word recognition accuracy for ratios automatically controlled by simple counting and topic modeling. The optimal fixed mixing ratio was what gave the maximum value of the red curve.

In the result of simple counting, the absolute difference of recognition accuracy was less than 1 point compared with the optimal fixed ratio for two (#2, #3) of the five dialect speakers (see Table 1(a)). This was true for four (except #2) of five CL speakers (see Table 1(b)). Note that the results for both dialect and CL speakers were produced with no different parameter settings; only the input utterances differed. The improvements over the recognition accuracies with the fixed mixing ratio of 100% (dialect) or 0% (CL) (without dialect mixing) were statistically significant at the $p = 0.05$ level by t-tests (dialect: $p = 0.017$, CL: $p = 0.011$).

Topic modeling showed higher recognition accuracy than simple counting for only two (#1, #2) of dialect speakers. The improvement over the recognition accuracy was statistically significant at the $p = 0.05$ level ($p = 0.027$) for dialect speakers, while not for CL speakers ($p = 0.78$). The followings are possible reasons topic modeling did not work well.

Error of prediction: Once recognition of words specific to a dialect fails, it affects the prediction more strongly than does simple counting, due to the difference among η_i. The value of parameter σ^2 was approximately 0.067; the prediction potentially includes an error of $\pm\sqrt{0.067} = \pm 0.26$.

Number of topics: To obtain a more accurate prediction model, we must choose the proper number of topics. This problem will be solved by non-parametric Bayesian modeling [21] that determines the proper number automatically.

6. Conclusion

We proposed and evaluated methods for controlling a dialect mixing ratio for recognizing utterances in dialects. Our method was twofold: simple counting and topic modeling. Either of them improved word recognition accuracy for both dialect and CL utterances. Simple counting showed high recognition accuracy for most speakers. Topic modeling surpassed simple counting for some speakers, but its parameters will need to be configured properly to obtain better recognition accuracy for all speakers. Recognizing multiple dialects will be our next step.

Future work includes extending the coverage of dialects geographically widely spread. We might have to model dialects hierarchically; (1) the probabilities of *pronunciations* are mixed for dialects in nearby areas, and (2) the probabilities of *sentences* are mixed for dialects in remote areas. Here, the restriction of unchanged word order is required for dialects in only nearby areas. Even hierarchical models, of course, will require appropriate dialect ratio to properly recognize sentences in any dialect. The result obtained in this paper will be the basis of dialect speech recognition for general purposes.

7. Acknowledgments

This study was partially supported by a Grant-in-Aid for Scientific Research (S) (No. 24220006).

8. References

[1] M. Anusuya and S. Katti, "Speech recognition by machine: A review," *International Journal of Computer Science and Information Security*, vol. 6, no. 3, pp. 181–205, 2009.

[2] N. Hirayama, S. Mori, and H. G. Okuno, "Statistical method of building dialect language models for ASR systems," in *Proc. of COLING 2012*, 2012, pp. 1179–1194.

[3] L. J. Brinton and M. Fee, *English in North America*, ser. The Cambridge history of the English language, J. Algeo, Ed. The Press Syndicate of the University of Cambridge, 2001, vol. 6.

[4] E. R. Thomas, *The Americas and the Caribbean*, ser. Varieties of English, E. W. Schneider, Ed. Mouton de Gruyter, 2008, vol. 2.

[5] D. Ramon, *We are one people separated by a common language*. iUniverse, 2006.

[6] H. Woods, "A socio-dialectology survey of the English spoken in Ottawa: A study of sociological and stylistic variation in Canadian English," Ph.D. dissertation, The University of British Columbia, 1979.

[7] D. Cruse, *Lexical Semantics*. Cambridge University Press, 1986.

[8] D. M. Blei and J. D. McAuliffe, "Supervised topic models," *arXiv preprint arXiv:1003.0783*, 2010.

[9] P. Ching, T. Lee, and E. Zee, "From phonology and acoustic properties to automatic recognition of Cantonese," in *Proc. of Speech, Image Processing and Neural Networks, 1994*, 1994, pp. 127–132.

[10] D. Miller and J. Trischitta, "Statistical dialect classification based on mean phonetic features," in *Proc. of ICSLP 1996*, vol. 4, 1996, pp. 2025–2027.

[11] W. Wolfram, *Ethnolinguistic Diversity and Literacy Education*. Routledge, 2009.

[12] D. Lyu, R. Lyu, Y. Chiang, and C. Hsu, "Speech recognition on code-switching among the Chinese dialects," in *Proc. of ICASSP 2006*, vol. 1, 2006, pp. 1105–1108.

[13] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *Proc. of CIAA 2007, Lecture Notes in Computer Science*, vol. 4783. Springer, 2007, pp. 11–23.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[15] National Institute for Japanese Language and Linguistics, Ed., *Database of Spoken Dialects all over Japan: Collection of Japanese Dialects (In Japanese)*. Kokushokankokai, 2001–2008, vol. 1–20.

[16] T. Misu and T. Kawahara, "A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts," in *Proc. of ICSLP 2006*, 2006, pp. 9–12.

[17] K. Maekawa, "Balanced corpus of contemporary written Japanese," in *Proc. of the 6th Workshop on Asian Language Resources*, 2008, pp. 101–102.

[18] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 7–12.

[19] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Acoustical Society of Japan (English Edition)*, vol. 20, pp. 199–206, 1999.

[20] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *Proc. of EuroSpeech 2001*, 2001, pp. 1691–1694.

[21] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *Journal of the ACM (JACM)*, vol. 57, no. 2, pp. 1–30, 2010.