Creating Dialect Sub-corpora by Clustering: a case in Japanese for an adaptive method

Yo Sato, Kevin Heffernan
Satoama Language Services, Kwansei Gakuin University
Kingston-upon-Thames, U.K., Sanda, Hyogo, Japan
yo@satoama.co.uk, kevin@kwansei.ac.uk

Abstract

We propose a pipeline through which to derive clusters of dialects, given a mixed corpus composed of different dialects, when their standard counterpart is sufficiently resourced. The test case is Japanese, where the written standard language is sufficiently equipped with adequate resources. Our method starts by detecting non-standard contents first, and then clusters what is deemed dialectal. We report the results on the clustering of mixed Twitter corpus into four dialects (Kansai, Tohoku, Chugoku and Kyushu).

Keywords: dialect, Japanese, clustering, adaptation of language model

1. Introduction

In what follows we propose a pipeline through which to derive clusters for dialects, given a body of 'mixed' corpus composed of different dialects, when one of them is sufficiently resourced. The main component of the proposal is an unsupervised clustering method to generate dialectal sub-corpora. The test case we experiment on is Japanese, where the variety deemed 'standard', the Tokyo dialect, is equipped with an adequate language model. We call such a dominant and well-resourced language a 'pivot' language, and its model a pivot model. Our main target data are Twitter utterances, coming from four broad dialect-speaking regions of Japan outside the Tokyo area. We will show how a) the dialectal content can be identified in this dataset and b) the identified dialectal content can be classified into the four dialects.

While our experiments are specifically on Japanese, we believe the work will have a wider implication, since a similar situation exists in many linguistic communities, where a well-resourced pivot exists but dialects lack mature enough language models. Ironically enough this is despite the fact that the availability of dialect data has increased in the cyberspace, with the rise of social media and interactive message boards. It is principally the lack of classified data that prevents the data from being utilised.

Japanese poses an additional challenge since there is no word segmentation in its orthography. This challenge however is a *general* problem. Word segmentation is an artefact of orthography, which is not present in speech, and hence requires lexical knowledge to perform. Thus the challenge is similar to the situation faced by a dialect monolingual who is exposed to other dialects in a *spoken* form: the difficulty is not so much with the unknown 'words' as the incomprehensible 'chunks' that may or may not correspond to words.

Our proposed procedure consists of three main stages. At the first stage we set apart the portion that is deviant in the view of the word pivot model. We find that this separated part not only contains dialects but non-dialectal non-standard utterances (e.g. internet jargon), so a second filtering follows, to set apart genuinely dialectal sentences. Here, in contrast to the prior stage, the *character* pivot model is used for filtering. The final stage consists of clustering, performed on what remains, where the technique employed is a modified form of divisive hierarchical clustering.

2. Related work

The present work may be categorised into the domain of 'discrimination of similar languages', which has attracted attention in recent years. Discrimination tasks have been tackled for variant sets such as South-Asian languages (Ranaivo-Malançon, 2006), English varieties (Lui and Cook, 2013) and Arabic (Zaidan and Callison-Burch, 2013). Work targeting dialects in the usual sense also exists, and Vergez-Couret and Urieli (2014) use what we call a pivot model in their study. Predominantly the techniques used in this domain are supervised, but Scherrer (2014) is a notable exception we find particularly relevant, in its unsupervised approach with pivot language models.

Another genre of studies from which we draw inspiration is unsupervised clustering that uses a distance metric. Nagata (2014) uses the distance *between* language models for clustering various types of English, though our notion pertains to the distance of a sentence *from* a language model. This notion, distance of a sentence from a certain model, has come to be used in corpus linguistics, where one might want to quantify the degree of difference from some existent model e.g. Collins et al. (2014) as a 'colloquiality' measure, Chen (2016) for proximity of textbooks to naturally occurring corpora.

Another line of research worth mentioning in the context of 'deviation from the standard' concerns normalisation. Research of this kind tries to deal with 'noisiness' by trying to find the 'standard' forms, faced with noisy corpora like Twitter. From a purely technical perspective our work also addresses the issue of 'deviation from the standard' faced with nonstandard data.

Like the present work Han et al. (2011) propose a 'pipeline' of three stages, and their initial stages also pertain to filtering non-standard content. With a similar goal (normalisation), Saito et al. (2014) introduce additionally a character-lattice based learning for non-segmented (Japanese Twitter) data, so this approach can be compared to ours, though theirs is a supervised method.

3. Data

We use two sets of data, controlled and natural. The first is the parallel dialect corpus recently published (Parallel Speech Corpora of Japanese Dialects, Yoshino et al. (2016) henceforth PCJD), with four sets of sentences that each represent a dialect (Tohoku, Kansai, Chugoku and Kyushu). Each set consists in turn of five sets of 100 sentences, the translations by five native speakers of the dialects, of the Tokyo dialect equivalents. Although the quantity of data is insufficient for training purposes, this dataset provides useful sources for evaluation.

Our main dataset is crawled Twitter data, obtained from the social media's public API (Twitter Inc., present) for the four-month period between February and May 2017, which amount to about 280 thousand sentences altogether after cleaning.[1] They were collected by setting geographical locations to the five areas represented in PCJD. Presumably due to the tweeting population difference, we have a slight disproportion: Tokyo roughly accounts for 31%, Kansai 28%, Kyushu 21%, Chugoku 16% and Tohoku 14%.

The data were then processed as follows. First, the whole corpora were processed with MeCab, a Japanese morphological analyser (Kudo et al., 2004). MeCab assigns each sentence a probability score, based on a CRF-trained model. With its training utility, we created our own model based on the written newspaper corpus (Kawahara et al., 2002; Mainichi Shimbun, 1995). We used the default dictionary that comes with the tool with nine features (Japan Information-Technology Promotion Agency, 1995). This constitutes our word pivot model, the basis for the first stage filtering.

Using roughly the same training method, we also built a pivot *character* model, this time using the Tokyo-area part of our crawled Twitter. Unlike the word model, of which it would be difficult to create reliable annotation, a character model can be built without much manual effort, and hence could be made easily available.

In short, we have made available the pivot models of standard Japanese in two types, word (newspaper) and character (Twitter). The target data is the Twitter data coming from four dialect-speaking regions. The goal is to create four sub-corpora of these dialects, by first detecting dialectal contents, and then, clustering them.

4. Method

In the following subsections we describe our three stages to separate and cluster dialectal utterances the Twitter data in more details.

4.1. Stage 1: Separation of non-conformant sentences

At this initial stage we separate the sentences which do not conform to the word pivot model. Since this CRF model outputs a probability given its dictionary items and their feature constellation in context, non-congruant items usually manifest themselves as their low probability, in one of the following two ways, faced with out-of-vocabulary (OOV) items. First, it could force some alternative word assignment in what could be called the forced-alignment strategy, finding an alternative segmentation path, at the likely cost of contextual probabilities. For example, for the sentence

そんなんありえへん家計ピンチやし

our word pivot model finds for the position of へん, in place of the correct label, dialectal negative auxiliary, an adjectival stem meaning 'strange'. It also labels やし, correctly two dialectal auxiliaries, as a noun (name of a fruit). As a result of forcing out-of-context words, the contextual probability suffers, in relation to our 'expected' probability.

Second, it might choose to, or be forced to, abandon the dictionary search, leaving the OOVs as unknown. In this case, the model invokes a user-defined OOV model. We adopt a scheme dependent primarily on the length,[2] giving more weights to the average length of (known) word tokens (approximately 3.2 syllables). We gave apriori initial values to OOV items, depending on its syntactic categories and adjusted experimentally, starting from the default values of the parameterised dictionary that came with the tool.

We use the concept of *distance from the pivot model* in this, as well as the following, stage as the filtering criteria, which is the difference in probability from the mean probability of the training data, that is, given the model $\mathbb{M}$ and the mean probability of the training data of $\mathbb{M}$, $E(P(s_{\mathbb{M}}))$

$$Dist(s, \mathbb{M}) = P(s)^{\mathbb{M}} - E(P(s_{\mathbb{M}})) \quad (1)$$

To determine what 'low' probability is low enough to be marked as non-conformant, we started from the mean distance we obtained for the Twitter corpus against the pivot model, and gradually lower it to optimise.

[1] We excluded utterances unsuitable for language model building, such as sentences that are heavily duplicated, are short (we set the minimum of 10 characters), consist only of punctuations, emojis and onomatopoeia.

[2] We also used the character type for assigning weights, given three types of alphabets (hiragana, katakana and kanji) in Japanese, preferring single-type sequences over mixed-type sequences.

With the threshold thus set, 41% of the data are classified as non-conformant overall.[3] For evaluation, with 100 sentences randomly extracted from the data coming from each region (i.e. 400 in total), we asked four native speakers, one each from our four regions, to check all the 400 sentences.[4] Each volunteer judged whether each sentence belongs to their dialect or not. The result is shown in the table below.

		Regional breakdown			
	All	Tohoku	Kansai	Chugoku	Kyushu
Utt. count	400	100	100	100	100
Matches	323	68	91	83	81
Other dialect	19	2	5	8	4
Nondialect	58	29	3	11	15
Precision	.808	.780	.940	.890	.910
Recall	.946	.994	.973	.978	.988

Table 1: Evaluation of non-conformity separation

The third row in the table shows two types of non-matches, ones belonging to one of our other three dialects and non-dialect. As can be seen, while we have achieved a reasonable success in setting apart the nonstandard portion, an obvious remaining problem is that this portion still contains a large proportion of non-dialect sentences. This is likely due to the fact that our pivot model is based on newspaper data. Therefore typical 'internet-speak' with jargon and contracted forms, such as 'ワロタ' or 'mjd', are frequently observed but mark unsurprisingly a low score, on par with the level of distinctly dialectal utterances.

4.2. Stage 2: Differentiating dialectal and non-dialectal deviations

At the second stage, we attempt to further filter the non-conformant sentences down to genuinely dialectal sentences. In order to do this, we use the same notion as the first stage, i.e. distance from the pivot model, but used the character model as a pivot. We also limit our target to substrings that are low in probability, or *anomalous segments* (ASs). By this we mean the segments, classified as either OOVs or known words, which are assigned a low probability. Again we adopted the simple threshold of the mean distance against the pivot model. The rationale behind this strategy is this: albeit both 'unknown' with respect to the pivot dictionary, the difference between dialectal and non-dialectal strings should manifest itself in the character-level patterns.

Our character pivot model is a three-dimensionality CRF model. The features used are the vowel/consonant types and 'voicedness', important elements that determine the character constellation of Japanese.[5] The distance from the pivot model is computed as in Equation 1, though now on the character model. We take the average probability of all the strings to be the 'expected' value, and consider the difference of the target substring to this expected value to be the distance from the pivot (character) model. The threshold for non-conformity has been determined experimentally.

Evaluation of the results at this stage has been done in two ways, without further assistance of our volunteers. First, to see its effect directly and allow comparison, we first inspected how many of the remaining (non-dialectal) sentences have been removed, with the same set of 400 sentences as before and recalculated the precision. The improvement figures are shown in Table 2. As can be seen, now approximately half of the non-dialectal sentences go away, bringing the precisions to a respectable level. Additionally, given the small size of this part of the results, one of the authors, a Tokyo dialect speaker, inspected results of 200 randomly picked sentences to see how many non-dialectal, or pivot dialect, sentences remain. We find 18, which is comparable to the results we obtained for the 400 set.

	Reduction nondialect	Improvement precision
Tohoku	$29 \rightarrow 19$	$0.68 \rightarrow 0.78$
Kansai	$3 \rightarrow 1$	$0.91 \rightarrow 0.98$
Chugoku	$11 \rightarrow 4$	$0.91 \rightarrow 0.95$
Kyushu	$15 \rightarrow 8$	$0.81 \rightarrow 0.91$
Overall	$58 \rightarrow 32$	$0.81 \rightarrow 0.91$

Table 2: Improvements at second filtering for dialect

4.3. Stage 3: Clustering

In our third and final stage, we cluster our likely 'dialectal' portion. We use a hierarchical clustering method, a top-down variety called DIANA (Kaufmann and Rousseeuw, 1990). We again require a character-based distance given the lack of dialect lexicon, though this time between sentences. The basis for our metric is Levenshtein distance (LD), which captures a major characteristic of the membership to the same dialect, i.e. sharing of non-standard lexical items, by restricting the target, as in 4.2., to the anomalous segments (ASs). We use two sets of ASs here, both character- and word-levels, in order to capture the anomaly of syntactic/contextual type on the first level, and that of phonotactic type on the second. Our *per sentence Levenshtein distance (LD_{sent})*, then, is as follows. Let AS_s and AS_l, be anomalous segment sets, then:

$$LD_{sent} = \sum_{i \in AS_s, j \in AS_l} \underset{i}{argmax} LD(i,j)/|AS_s| \quad (2)$$

[3] There was a significant difference in terms of the proportion of non-conformity across regions, however, with 34% for Tohoku, 52% for Kansai, 45% for Kyushu and 40% for Chugoku. See Sato and Heffernan (2017) for details.

[4] Later another native speaker of each region was asked to validate the prior volunteer's judgements and excluded the ones with disagreement.

[5] The surface N-gram model does not suffice for Japanese because Japanese characters mostly consist of a combination of a consonant and a vowel and surface forms may obscure a similarity in sound between two characters.

	Tohoku	Kansai	Chugoku	Kyushu	Recall
Tohoku	55	2	5	3	.846
Kansai	3	67	8	6	.798
Chugoku	2	10	59	11	.719
Kyushu	9	7	8	57	.703
Precision	.797	.779	.738	.740	

	Tohoku	Kansai	Chugoku	Kyushu	Recall
Tohoku	86	2	6	5	.868
Kansai	0	68	13	9	.756
Chugoku	9	5	66	12	.717
Kyushu	4	8	9	73	.776
Precision	.869	.819	.702	.759	

Table 3: Clustering performance, Twitter and PCJD

which is essentially the average of the LDs between the segments in the smaller set and their closest counterparts in the larger set.

However, this proves insufficient due to the fact that there are lexical items which, while not found in the pivot lexicon, belong to more than one dialect, particularly (not surprisingly) between neighbouring dialects. For example, the auxiliary よる and じゃ are both used frequently in the Chugoku and Kyushu dialects. Therefore, an entirely bottom-up procedure of clustering (such as agglomerative clustering) could lead to wrong groupings at the initial stages and cannot recover from these mistakes.

We have therefore opted to have a top-down constraint incorporated, the *sequence sharing rate* or SSR. The intuition is that a dialect will have a *consistent* shared vocabulary, and hence, even if some words can happen to be shared across dialects, the substring sharing *as a whole* inside a dialect should be higher than across dialects. By shared sequence we mean a contiguous substring that is found in the target strings. We take the longest match. Therefore for example between *abcde* and *ijbcdk* it is *bcd*. We also take multiple matches if they exist but not repetitions in the same string, so for *abcdef* and *efabcef* we will have two shared sequences, *ef* and *abc*. Given a set of utterances U and a set of shared sequences that a set of shared sequences S, SSR is defined as follows:

$$SSR(U) = \sum_{s \in S} (len(s) \times 2)/|U| \quad (3)$$

where $len(s)$ refers to the number of characters in shared sequence s. Notice we give more weights, proportionate to two, to longer shared sequences, given the likelihood that longer sequences contain words and phrases, which we are implicitly modelling.

Now, DIANA essentially splits the chosen cluster into two subsets iteratively. The choice of which cluster to be split is made between iterations.[6] The binary split proceeds in such a way that the original cluster, below A, 'transfers' its 'most dissimilar' of the remaining members to a new cluster, B, with the following function for *dissimilarity distance*:

$$D_{dissim} = \frac{1}{|A|-1} \sum_{j \in A} d(i,j) - \frac{1}{|B|} \sum_{k \in B} d(i,k) \quad (4)$$

where $i, j \in A$ and $k \in B$ and d is the distance metric. $argmax_i D_{dissim}$ is the most dissimilar member to transfer, and this continues until $D_{dissim} \leq 0$.

With per-sentence Levenshtein distance as the distance metric, an iteration proceeds generally as described above. The split is then adjusted after each iteration in such a way that the two metrics, LD and SSR, are balanced. We do this by reversing some of the decisions made in the iteration. We start from the least dissimilar, or the most recent, item 'moved' from the original cluster to the new, and conduct a check towards the less dissimilar ones iteratively. The check is about whether taking the item back to the original cluster improves the SSR. If so, the previous decision is reversed and the item goes back to the original. We repeat this procedure until the point where no further improvement is likely.

In Table 3 we report the performances, for the Twitter data (filtered by our native volunteers to what they judged dialectal as the Gold Standard) and PCJD, in the form of confusion matrices. For PCJD, we used a set of 100 sentences for each of the four dialects, though we excluded a few sentences[7]. For this dataset we obtained on average precision and recall around 80% level, but have a large variance between dialects. In particular precision remains rather low for Chugoku and Kyushu, presumably due to the confusability between the dialects. The Twitter results follow roughly the same pattern, with a slightly lower average generally than PCJD. The aforementioned inconsistency between regions is less pronounced however. This can be due to the effectiveness of the top-down control, which can only kick in when the data reaches a critical mass.

5. Final remarks and future tasks

We have presented a pipeline of methods that generates clusters of dialects from a mixed corpus, on the basis of a pivot language model.

As has been said in Introduction, our method is general enough to be replicated with other languages. Although we did not emphasise it (since no effective test has been feasible), the fact that there is no restriction on the number of clusters is a great advantage for generality.

[6] Amonst the several criteria for calculating the distance, we use the 'average linkage' criterion.

[7] We excluded 19 sentences altogether out of total 400, for the following two reasons: a) ones that are close to the pivot model, and hence would have been isolated in our step 2, b) ones that are identical between regions.

A shortcoming on the other hand, related to generality of application too, is that it depends heavily on the pivot model. In our case, the pivot model was a CRF-based model, which in turn depends heavily on feature configurations. We also have had to resort to some feature engineering to make clustering work. There will likely be a trade-off between performance and amount of such feature manipulations. Future research therefore should address a more general training method.

Another major issue is evaluation, which can be labour-intensive if manually done, like we did. A small test data also means less reliability and generality in the results. It would therefore be desirable to handle Gold Standard creation more efficiently, e.g. through crowd sourcing, or devise a way for intrinsic evaluation.

The most important further goal we envisage is to create automatically the language model for each new subcorpus. We believe that as long as lexicons are similar, it is possible to Scherrer (2014)'s method, i.e. generating sub-corpus lexicons by finding equivalent words. This will enhance greatly the usefulness of the dialect corpora, and also render intrinsic evaluation feasible.

6. Bibliographical References

Chen, A. C.-H. (2016). A critical evaluation of text difficulty development in elt textbook series: A corpus-based approach using variability neighbor clustering. system. *System*.

Collins, P. A. M., Borlongan, and Yao, X. (2014). Modality in philippine English: A diachronic study. *Journal of English Linguistics*.

Han, B., Cook, P., and Baldwin, T. (2011). Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 49th Annual Meeting of ACL*.

Kaufmann, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data - An Introduction to Cluster Analysis*. John Wiley and Sons.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying Conditional Random Field to Japanese morphological analysis. In *Proceedings of the conference on Empirical Method in Natural Language Processing*.

Lui, M. and Cook, P. (2013). Classifying English documents by national dialect. In *InProceedings of Australasian Language Tchnology Workshop*.

Nagata, R. (2014). Language family relationship preserved in non-native English. In *Proceedings of COLING 2014*.

Ranaivo-Malançon, B. (2006). Automatic identification of close languages - case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology*.

Saito, I., Kugatsu, S., Asano, H., and Matsuo, Y. (2014). Morphological analysis for Japanese noisy text based on character-level and word-level normalization. In *Proceedings of COLING 2014*.

Sato, Y. and Heffernan, K. (2017). Quantifying 'standardness' of the language use in locality: a study with twitter data. In *Extended abstract, Corpus Linguistics International Conference 2017*.

Scherrer, Y. (2014). Unsupervised adaptation of supervised part-of-speech taggers for closely related languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*.

Vergez-Couret, M. and Urieli, A. (2014). Pos-tagging different varieties of occitan with single-dialect resources. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*.

Yoshino, K., Hirayama, N., Mori, S., Takahashi, F., Itoyama, K., and Okuno, H. G. (2016). Parallel speech corpora of Japanese dialects. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages –, Paris, France, may. European Language Resources Association (ELRA).

Zaidan, O. F. and Callison-Burch, C. (2013). Arabic dialect identification. *Computational Linguistics*.

7. Language Resource References

Japan Information-Technology Promotion Agency. (1995). *IPA dictionary*. Japan Information Technology Promotion Agency.

Daisuke Kawahara and Sadao Kurohashi and Koiti Hasida. (2002). *Kyoto University Text Corpus*. Kyoto University.

Mainichi Shimbun. (1995). *Mainichi Newspaper CD-ROM 1995*. Mainichi Newspaper.

Twitter Inc. (present). *Twitter Stream API*. Twitter Inc.