Quantifying 'standardness' of the language use in a locality: a study with Twitter data

Yo Sato[1], Kevin Heffernan[2], Shunsuke Kishie[3], and Kota Hattori[3]

[1]Satoama Language Services, London, U.K.

[2]Kwansai Gakuin University, Sanda, Japan

[3]Tokushima Univerisity, Tokushima, Japan

In this work we present a probabilistic method to quantify the distance of a corpus from the standard model. We then apply this metric to show that the degree to which the actual language use deviates from the standard language varies significantly from region to region. Specifically, with the Twitter data coming from four broad regions in Japan, Tohoku, Kansai, Chugoku and Kyushu (about 17,000 tweets each), we observe widely differing distances, which indicates that the readiness to conform to the standard language varies between regions.

We started by training a model with a newspaper corpus written in standard Japanese and a lexicon of standard Japanese. For this training, as well as for computing the plausibility of a sentence, we use the morphological analysis tool MeCab (Kudo, Yamamoto, & Matsumoto, 2004). This tool employs the Conditional Random Fields (CRF) algorithm (Lafferty, McCallum, & Pereira, 2001) to build the feature-based statistical model, and computes the 'cost' based on the probability of a sentence (the smaller the cost the greater the conformity to the model). CRF, in its training, takes into account not just the words' surface forms but their various attributes, as well as their order. We used in our training nine features including lemma and part of speech. MeCab further considers multiple paths for different word segmentation hypotheses. This functionality of MeCab, designed primarily to adapt to the texts in an orthography with no space between words, nicely mimics the real situation in speech where no apparent word segmentation is given. Thus, we are closely modelling the situation where a nono-dialectal speaker is exposed to a non-standard dialectal sentence, which may pose him/her a variety of challenges related to either grammar, vocabulary or word segmentation.

We then confirmed the expectation that on average, greater costs are in fact associated with dialects, using pre-classified parallel dialect corpora developed by (Yoshino et al., 2016). The corpora consist of a set of standard Japanese sentences as well as its translations into the dialects of the above four regions, all hand-crafted to ensure that there will be great contrasts between dialects. Reflecting this intended contrast, the average cost for the standard-Japanese is $-1,570$, while that for each dialect is much higher (over 30,000).

Taking these numbers as expected costs, we then processed our region-classified Twitter data and compared their costs. The distance for each region is calculated as the expected cost minus the twitter cost, producing the following figures:

Tohoku $56,148 - 30,301 = 25,847$
Kansai $31,646 - 32,279 = -633$
Chugoku $32,477 - 31,876 = 571$
Kyushu $39,649 - 31,012 = 8,637$

Thus the results show that for our four regions, the size of deviation from the standard language varies from one another. In particular, Kansai dialect speakers seem the most willing to adhere to their own dialect, while Tohoku dialect speakers are the least willing. This confirms the conventional observation by the Japanese people.

We also compared the differences in cost of four regions inside the twitter corpora against the standard-Japanese speaking region (Kanto, at 29,747). The ANOVA analysis shows that the regional deviations from this figure are mostly significant ($P < .01$) with the exception of Tohoku dialect speakers, reinforcing the aforementioned contrast in regional characteristics in terms of willingness/reluctance to speak their dialect.

Our method of measuring distances between subgroups of a corpus is a general one, in the sense that it can be straightforwardly extended to other languages to test their diversity, as long as there is a 'standard' set of resources (lexicon and annotated corpus). Also, the target parameter does not have to be dialect, and could be of any genre: one could for example apply the method to see if the language use of a social subgroup diverges from the society as a whole.

The main limitation of our method on the other hand is that it only shows the divergence from a certain standard, not the differences between any given set of groups. Also, there is no guarantee either that the distance we observed is actually due to a dialectal difference (if likely). To be able to distinguish between any regions, i.e. cluster dialects, we plan therefore to complement it with other metrics, such as vocabulary share rate (as in (Inoue, 2008)).

References

Inoue, F. (2008). Geographical distance center and multivariate analysis of the standard japanese. Dialectologia(1), 65–81.

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying Conditional Random Field to Japanese morphological analysis. In Proceedings of the conference on empirical method in natural language processing.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the eighteenth international conference on machine learning (pp. 282–289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=645530.655813

Yoshino, K., Hirayama, N., Mori, S., Takahashi, F., Itoyama, K., & Okuno, H. G. (2016, may). Parallel speech corpora of japanese dialects. In N. C. C. Chair) et al. (Eds.), Proceedings of the tenth international conference on language resources and evaluation (lrec 2016) (pp. –). Paris, France: European Language Resources Association (ELRA).