

Automatic Method to Build a Dictionary for Class-based Translation Systems

Kohichi Takai¹, Gen Hattori¹, Keiji Yasuda¹, Panikos Heracleous¹, Akio Ishikawa¹, Kazunori Matsumoto¹ and Fumiaki Sugaya^{1,2}

¹ KDDI Research, Inc.

Garden Air Tower, 3-10-10, Iidabashi, Chiyoda-ku, Tokyo, 102-8460, Japan

{ko-takai, ge-hattori, ke-yasuda,
pa-heracleous, ao-ishikawa, matsu}@kddi-research.jp

² MINDWORD Inc.

7-19-11 Nishishinjuku, Shinjuku-ku, Tokyo, 160-0023, Japan

fsugaya@mindword.jp

Abstract. Mis-translation or dropping of proper nouns reduces the quality of machine translation or speech translation output. In this paper, we propose a method to build a proper noun dictionary for the systems which use class-based language models. The method consists of two parts: training data building part and word classifier training part. The first part uses bilingual corpus which contain proper nouns. For each proper noun, the first part finds out the class which gives the highest sentence-level automatic evaluation score. The second part trains CNN-based word class classifier by using the training data yielded by the first step. The training data consists of source language sentences with proper nouns and the proper nouns' classes which give the highest scores. The CNN is trained to predict the proper noun class given the source side sentence. Although, the proposed method does not require the manually annotated training data at all, the experimental results on a statistical machine translation system show that the dictionary made by the proposed method achieves comparable performance to the manually annotated dictionary.

1 Introduction

As a result of drastic advances in technical innovations of speech processing and natural language processing, a speech-to-speech translation system is becoming a realistic tool for travelers. Especially for the travel domain, the coverage of proper nouns for tourist spots, landmarks, restaurants, and accommodations highly influences system performance.

Okuma et al. [1] proposed a class based method to install hand-crafted bilingual dictionaries into a Statistical Machine Translation (SMT) framework to improve the proper noun coverage of Machine Translation (MT). However, there are remaining problems for practical usage. It is the cost of building the dictionary. Since the number of proper noun such as restaurant names or product

names, is increasing daily, the cost to manually maintain a dictionary is very high. Especially for the speech translation system using a class-based language model, there is an additional cost to annotate word class for each new word. In this paper, we propose a method to automatically estimate word class.

Section 2 describes related work in this research. Section 3 explains the proposed method. Section 4 shows the experimental results using the SMT-based system. Section 5 concludes the paper and presents some directions for future work.

2 Related Work

Class-based language model had been proposed in the automatic speech recognition research field to solve the problem of data sparseness. And, the idea also has been applied to SMT framework [1]. To apply class based idea to SMT framework, we have to maintain dictionary which includes translation pairs of words and their word classes. To automatically acquire translation pairs, several methods have been proposed, such as bilingual dictionary [2] and transliteration approach [3–5].

In this paper, we only focus on the word class annotation part. There are also many researches dealing with automatic annotation of word category, in named entity recognition and machine translation research field [6, 7]. These methods require manually annotated training data to train a classifier. The proposed method in this paper also uses supervised training approach. However, the proposed method automatically builds a training data by using bilingual corpus and class-based machine translation system which is the target system to use the dictionary. Although the bilingual corpus is required to build the training data, only monolingual corpus is required to actual word class classification.

3 Proposed Method

As shown in Fig. 1, the proposed method consists of training data building part and word classifier training part. Details of these two parts are explained in this section.

3.1 Training data building without human annotation

Fig. 2 shows the flow of training data building part. Broken lines in the figure indicate data flow. The first part uses bilingual corpus which contains proper noun. For each bilingual sentence pairs with proper nouns, the proposed method yields multiple machine translation results by the following steps.

Step1 Translation pairs of proper noun are extracted from bilingual sentence pairs from the corpus.

Step2 The extracted proper noun pairs are registered to the dictionary as one of word classes.

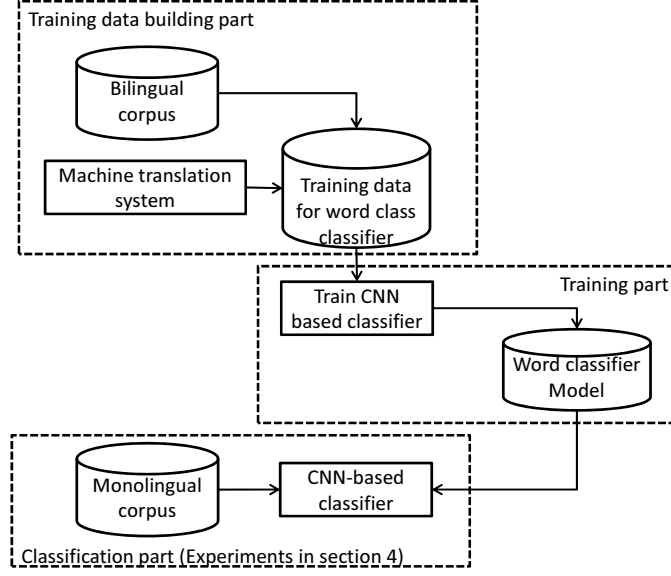


Fig. 1. Framework for the proposed method.

Step3 Translate the source side sentence containing the target proper noun using MT with the dictionary made in step2.

Step4 The registered proper noun entry is removed from the dictionary.

Step5 Step 3 to 4 are repeated until all classes are registered to the dictionary.

By using the multiple translations and reference sentence which is the target language side of the sentence translation pair, the most suitable word class (\hat{c}) for the target proper noun is chosen using the following formula.

$$\hat{c} = \operatorname{argmax}_{c \in C} S_{RIBES}(T_{REF}, T_{MT}^c) \quad (1)$$

where C , T_{REF} and T_{MT}^c are the set of word class, reference translation which is the target side sentence from bilingual corpus, and a translation result by MT whose dictionary has entry of target proper noun as word class c , respectively. And S_{RIBES} is the RIBES [8] score between T_{REF} and T_{MT}^c , calculated using the following formula.

$$S_{RIBES}(T_{REF}, T_{MT}^c) = R_{cor}(T_{REF}, T_{MT}^c) \times (l_{com}/l_{MT})^\alpha \quad (2)$$

where l_{MT} , l_{com} and R_{cor} are the total words number in T_{MT}^c , the number of common words between T_{REF} and T_{MT}^c and the rank correlation coefficient between the common words, respectively. $\alpha(0 \leq \alpha \leq 1)$ is the hyper parameter for penalty.

In the word classifier training part, pairs of source side sentence and \hat{c} which is teacher signal are used for training.

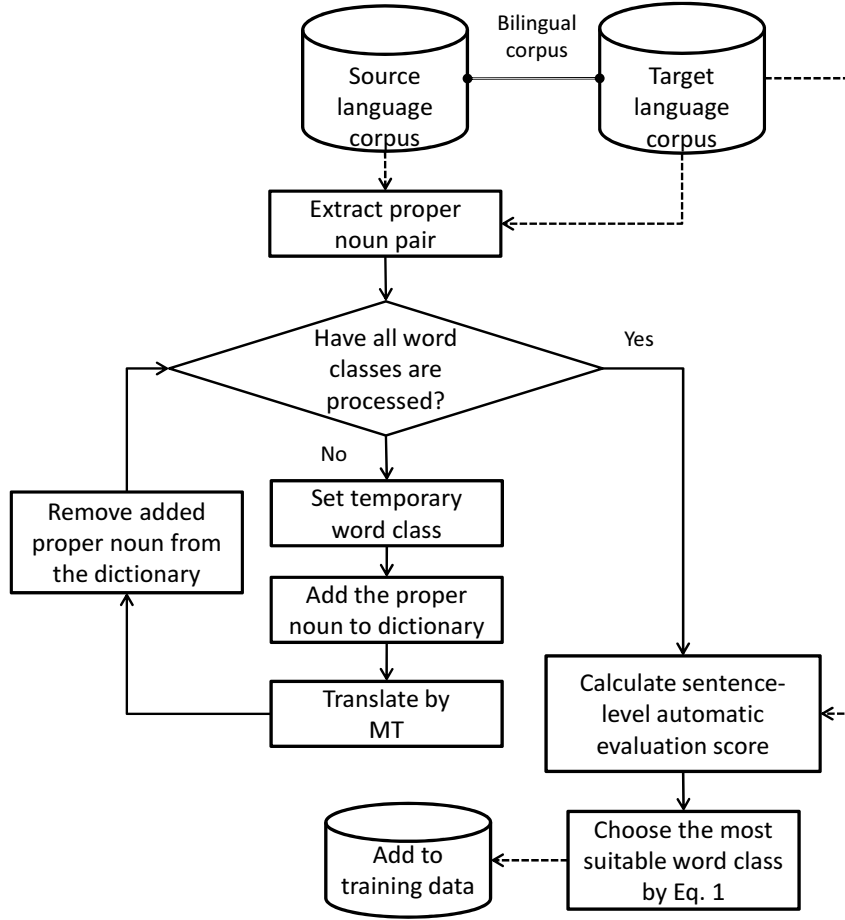


Fig. 2. Flow of the traning data building part.

3.2 Word classifier training part

This subsection explains the method to train word classifier using the training set build by the previous subsection.

For proper noun classification, we train a classifier which is configured as a Convolutional Neural Network (CNN). CNN showed superior performance in the research fields of image processing and speech recognition [9, 10]. Currently, CNN has also outperformed the Natural Language Processing (NLP) task such as text classification [11, 12] by incorporating word-embedding [13] in the input layer.

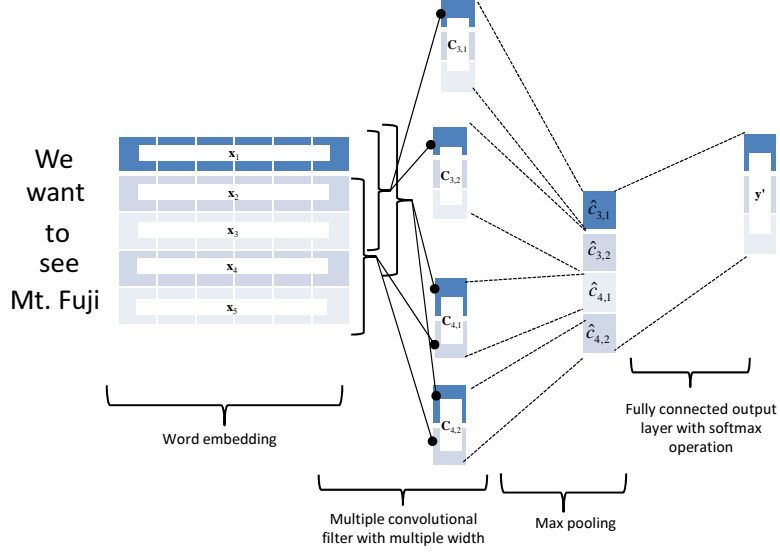


Fig. 3. Convolutional neural network for word classification.

The network configuration for our word classification is shown in Fig.3³. Here, $\mathbf{x}_i (\in \mathbb{R}^k)$ is the word-embedding vector of i -th word in a given sentence. By concatenating word-embedding vectors, a sentence whose length is n is expressed as follows:

$$\mathbf{x}_{1:n} = \mathbf{x}_1 \cdots \oplus \mathbf{x}_i \cdots \oplus \mathbf{x}_n \quad (3)$$

The convolutional layer maps the n -gram features whose length (or filter window size) is h to the j -th feature map by using the following formula:

$$c_{h,j,i} = \tanh(\mathbf{w}_{h,j} \cdot \mathbf{x}_{i:i+h-1} + b_{h,j}) \quad (4)$$

where $w_{h,j}$ and $b_{h,j}$ are weights for filtering and bias terms, respectively. For each n -gram length and feature map number, concatenate the results of Eq.4 as follows

$$\mathbf{c}_{h,j} = [c_{h,j,1}, c_{h,j,2}, \cdots, c_{h,j,n-h+1}] \quad (5)$$

The max pooling layer chooses the element that has the largest value from all elements in $\mathbf{c}_{h,j}$ as follows

$$\hat{c}_{h,j} = \max_{i=1}^{n-h+1} \mathbf{c}_{h,j} \quad (6)$$

Fully connected output layer takes the softmax operation to yield the probability distribution of the word class ($\hat{\mathbf{y}} \in \mathbb{R}^{n_c}$) as follows

³ The actual hyper parameters' setting is different from the example shown in the figure. Detail setting will be explained in section 4

$$\hat{\mathbf{y}} = \frac{\exp(z_q)}{\sum_{p=1}^{n_c} \exp(z_p)}, q = 1, \dots, n_c \quad (7)$$

where n_c is the total number of word classes. And, \mathbf{z} ($\in \mathbb{R}^{n_c}$) are the raw output values from output layer.

Table 1. Detail of word class in the data set.

Word class	% in the data set
Accommodation	10.33
Attraction	4.90
Building	8.44
Country name	12.15
Foreign First Name	5.57
Foreign Last Name	3.73
Food	7.84
Japanese First Name	4.35
Japanese Last Name	4.17
Land Mark	11.73
Organization	9.29
Shop	4.75
Souvenir	6.54
Others	6.20
Total	100

Table 2. Statistic of training corpora used for the experiments.

Corpus type	# of words	Lexicon size
BTEC (Japanese side)	83,942	9,266
Wikipedia corpus	10,363,151	116,556

4 Experiments

4.1 Evaluation Method

In the experiments, firstly, we build the training data and train the CNN-based classifier by the proposed method. Then, we build Japanese-English bilingual proper noun dictionary by using the classifier. To evaluate dictionary quality, we compare Japanese-to-English translation quality of SMT-based system with several conditions as follows.

Table 3. CNN parameter setting

Parameters	Setting
Maximum length of input sentence	150 words
Mini batch size	65
Dimension of word-embedding vector (k)	400
Filter window size (n -gram length)	3 to 5-gram
Number of filters for each window size	128
Drop out rate for fully connected layer	0.5
Optimizer	Adam optimizer
# of output units	14

Condition 1 Manual word class annotation by a human annotator.

Condition 2 Random annotation with uniform distribution.

Condition 3 Random annotation with prior distribution shown in Table 1.

Condition 4 Automatic word class annotation by the propose method.

We also use RIBES [8] for automatic evaluation of the translation quality. And, the evaluation experiments are carried out by 10-class validation manner.

4.2 Experimental Settings

For the extraction of proper nouns, we use Japanese and English parts of Basic Travel Expression Corpus (BTEC) [14]. By using an in-house dictionary, we extract 5,471 sentence translation pairs containing proper noun. Here, we extract sentences which only contain one proper in each sentence. Table 1 shows the details of the word class specification and occurrence in the training set. There are 14 word classes which are related to travel domain. In the data set, word class of country name has the highest occurrence. Meanwhile, foreign last name has the lowest occurrence.

Beforehand of training CNN-based classifier, we trained the word-embedding matrix using Word2Vec[13] on the Wikipedia corpus. The word-embedding matrix is fixed through CNN training. The detail of these two corpora are shown in Table 2.

Table 3 shows the detailed parameter setting of the CNN-based classifier. As shown in the table, the CNN has 14 output units, and each of them outputs the probabilities of the word classes, which are shown in Table 1. As shown in Table 1, data size varies depending on the word class. To reduce the adverse effect of data imbalances, we sampled training data to be balanced while mini batch training.

4.3 Experimental Results

Figure 4 shows the automatic evaluation results of translation by SMT with several kinds of class-based bilingual dictionaries. In the figure, the vertical axis

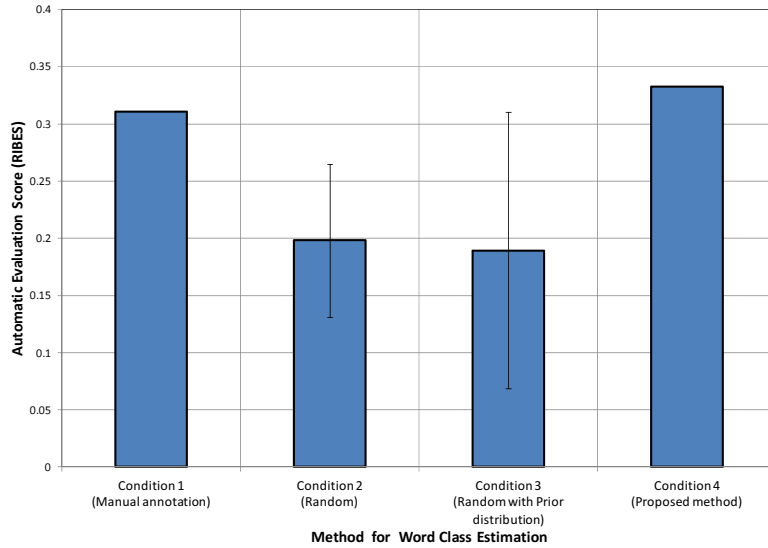


Fig. 4. Translation quality by SMT with several dictionaries.

represents automatic evaluation score calculated by RIBES. For all conditions, we use the same class-based SMT system proposed by [1] and same proper noun coverage. Difference is only the way to annotate word class for the proper noun dictionary. For condition 2 and 3, we carried out 10 times random annotations for each proper noun, then calculate an average score. The error bars in the figure show standard deviations over 10 times random trials. Meanwhile, for condition 4, we carried out 10-cross validation once on the data set.

As shown in the figure, the proposed method gives way better performance comparing to two random annotation. Additionally, the proposed method gives better performance than the manual annotation. The reason may be as follows. A human annotator decided word class from the proper nouns only. Meanwhile, the proposed method automatically annotates word class using the source side sentence including the proper noun, thus, the proposed method can use context information to annotate the proper noun. Such context information gives advantage for proper nouns which have multiple meanings.

5 Conclusions and Future work

We proposed a method to build a dictionary for a class-based translation system.

We carried out experiments using BTEC corpus on SMT. Firstly, training set for CNN-based word classifier was automatically made. Secondly, the CNN-based word classifier was trained using the training set. Then, we automatically labeled word class for the proper nouns in the test set and added to the dictionary for SMT. Finally, we translate the test set by SMT using the dictionary to

evaluate the dictionary quality by translation quality. According to the experimental results, the dictionary build by the proposed method have performance comparable to the manually annotated dictionary.

Since the proposed method does not require manual annotation, it enables quick development of class-based machine translation system.

As future work, we will increase the corpus size for improvement of classification accuracy. Although the experiments shown in the paper were carried out on the existing BTEC corpus, now we are carrying out speech translation field experiments to evaluate the effectiveness of the proposed method in the real field.

Acknowledgments

This research is supported by Japanese Ministry of Internal Affairs and Communications as a Global Communication Project.

References

1. Okuma, H., Yamamoto, H., Sumita, E.: Introducing a translation dictionary into phrase-based smt. The IEICE Transactions on Information and Systems **91-D** (2008) 2051–2057
2. Tonoike, M., Kida, M., Takagi, T., Sasaki, Y., Utsuro, T., Sato, S.: Translation Estimation for Technical Terms using Corpus collected from the Web. In: Proceedings of the Pacific Association for Computational Linguistics. (2005) 325–331
3. Al-Onaizan, Y., Knight, K.: Translating named entities using monolingual and bilingual resources. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). (2002) 400–408
4. Sato, S.: Web-Based Transliteration of Person Names. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. (2009) 273–278
5. Finch, A., Dixon, P., Sumita, E.: Integrating a joint source channel model into a phrase-based transliteration system. In: Proceedings of NEWS2011. (2011) 23–27
6. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics (2016) 1064–1074
7. Yasuda, K., Heracleous, P., Ishikawa, A., Hashimoto, M., Matsumoto, K., Sugaya, F.: Building a location dependent dictionary for speech translation systems. In: 18th International Conference on Computational Linguistics and Intelligent Text Processing. (2017)
8. Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H.: Automatic evaluation of translation quality for distant language pairs. In: Conference on Empirical Methods in Natural Language Processing. (2010) 944–952
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, Curran Associates, Inc. (2012) 1097–1105
10. Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **22** (2014) 1533–1545

11. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. (2014) 1746–1751
12. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: Convolutional neural networks for modeling sentences. In: Proceedings of the 52nd Annual Meeting for Computational Linguistics. (2014) 655–665
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26. Curran Associates, Inc. (2013) 3111–3119
14. Kikui, G., Sumita, E., Takezawa, T., Yamamoto, S.: Creating corpora for speech-to-speech translation. In: 8th European Conference on Speech Communication and Technology (EUROSPEECH). (2003) 381–382