Name Translation based on Fine-grained Named Entity Recognition in a Single Language

Kugatsu Sadamitsu, Itsumi Saito, Taichi Katayama, Hisako Asano and Yoshihiro Matsuo
NTT Media Intelligence Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847, Japan
{sadamitsu.kugatsu, saito.itsumi, katayama.taichi,
asano.hisako, matsuo.yoshihiro} @lab.ntt.co.jp

Abstract

We propose named entity abstraction methods with fine-grained named entity labels for improving statistical machine translation (SMT). The methods are based on a bilingual named entity recognizer that uses a monolingual named entity recognizer with transliteration. Through experiments, we demonstrate that incorporating fine-grained named entities into statistical machine translation improves the accuracy of SMT with more adequate granularity compared with the standard SMT, which is a non-named entity abstraction method.

Keywords: Statistical Machine Translation, Extended Named Entity, Bilingual Named Entity Recognition

1. Introduction

One of the issues in statistical machine translation (SMT) systems is named entity (NE) translation. There are two major problems with NEs. One, NEs cause data sparseness in training data due to the large range of NE tokens, and two, NEs are often treated as unknown words in test data because they have not been included in training data. For example, the two phrases "*went to Tokyo by train*" and "*went to Osaka by train*" are similar, but most translation phrase tables would not treat these as a similar pair because of the difference between "*Tokyo*" and "*Osaka*". Even worse, if only "*Tokyo*" appears in the training corpus, "*Osaka*" becomes a unknown word. One solution for these problems is abstracting NE for SMT and transliterating the NEs. Several approaches to this effect have been proposed (Li et al., 2013; Hermjakob et al., 2008; Hassan et al., 2007; Al-Onaizan and Knight, 2002; Knight and Graehl, 1997), but they used coarse NE classes such as "PERSON" and "GPE" and did not focus on the relevance between the granularity of NE classes and SMT. Dynamic clustering for chunks as EM algorithms (Och, 1999), in contrast, does not depend on the use of NEs. This method is effective in cases where the amount of the parallel corpus is enough, but it is reasonable to utilize another informative resource when the amount of a parallel corpus is short.

Our hypothesis is that focusing more on adequate fine-grained NE classes and incorporating them into SMT should improve the accuracy of SMT. In this paper, we propose an NE abstraction method with extended NE (ENE) labels (Sekine, 2008) as fine-grained NE labels and compare them with a standard SMT (with neither NER nor NE abstraction) using annotated bilingual NE test data.

To this end, we also have to construct high accuracy bilingual NER to abstract NE chunks in training bilingual corpus to NE labels. It is relatively easy to extract NEs from both source and target languages when the NERs or NE resources of both languages exist. However, occasionally, only fine-grained NE resources are available in a single language, as is the case with the Japanese NE annotated corpus belonging to Sekine's ENE definition including the 200 NE classes that we used here (Hashimoto and Nakamura, 2010). To overcome this issue, we constructed a bilingual NER that needs only a monolingual NER.

In this paper, although we use Japanese-English translation pairs, our methods and analysis are not language dependent and are applicable for other language pairs with NE resources.

2. Name translation based on fine-grained named entity recognition

2.1. Utilization of fine-grained named entity

Several studies on name translation using monolingual NER have been attempted, but it does not work. The typical problems associated with monolingual NER are presented as below (Hermjakob et al., 2008).

- Automatic named-entity identification makes errors.
- Not all named entities should be transliterated.

We consider that one factor of these disadvantages is granularity of NE classes, since coarse NE labels are

Label	NE num.	NE ratio (/LOC)	Label	NE num.	NE ratio (/LOC)
Country	1057	56.13	School	11	0.58
City	310	16.46	Island	10	0.53
Province	289	15.35	Amusement park	6	0.32
Domestic region	58	3.08	Airport	6	0.32
Continental region	32	1.70	Sports facility	5	0.27
Mountain	30	1.59	Theater	5	0.27
Worship place	15	0.80	River	4	0.21
Sea	13	0.69	Railroad	3	0.16
Station	13	0.69	Bridge	2	0.11
Museum	11	0.58	Zoo	2	0.11

Table 1: NE occurrence in 7,000 sentences (Hiragana TIMES).

mixed multiple types of fine-grain NE classes that are hard to recognize and transliterate. Therefore, we focus on a few fine-grain NE classes in accordance with importance and accuracies of the NE classes for target corpus.

2.2. The target classes of fine-grained named entities

In this paper, we utilized Sekine's ENE definition (Sekine, 2008) as fine-grained labels from two perspectives.

- It widely covers many types of NEs for adaptation of SMT.
- The monolingual (Japanese) training data is well developed (Hashimoto and Nakamura, 2010).

For selecting the target classes, we focus on the "Person" class and some sub-classes of "Location" because of the high possibility of transliteration. Other NE classes, "Artifact" and "Organization", are often inappropriately transliterated as "The association of natural language processing" and "言語処理学会 (pronounced as *gengo shori gakkai*)".
Table 1 shows the frequency of NE in 7,000 Japanese sentences in Hiragana TIMES using an automatic Japanese NER. In this table, "Country", "City", "Province", and "Domestic region" account for a high percentage. In these classes, we also confirmed that monolingual NER achieved a high accuracy over 85 in F value compared to the other classes (e.g., in the "Station" and "Museum" classes, NER only achieved 76 and 56 in F value, respectively). We removed "Country" so as to target just three labels, as "Country" is not applicable to transliteration ("日本 (nihon)" and "Japan") and we assume that NEs such as "Japan" and "U.S." occur often enough for training without NE abstraction.
In the experimental section, we examined two ways for treating the three labels, specifically, how well they kept their fine granularity and merged as one "sub-Location". Fine granularity can be used to make sophisticated translation models, although the accuracy of the fine granularity NER will be lower than that of the merged grain.

2.3. Our proposed bilingual NER and SMT systems

In this section, we introduce a system for bilingual NER and our SMT systems. To apply our systems even when we have only a monolingual NER trained by monolingual NE annotated corpus, we constructed a bilingual NER in a training step. This can be done using word alignment, translation dictionaries, phonetic similarity using transliteration, etc., but here, we focus on phonetic similarity only, for three reasons.

- We want to focus on the phrases that are possible to transliterate.
- When we utilize both NEs in source language sentence recognized by monolingual NER and automatically esitimated word alignment, the words in the target language words aligned from source language NEs often include the surrounding context in error, not just the NE chunk in target language. These errors lead to the difficulty of training accurate SMT models with abstracted NEs.
- Although utilizing translation dictionaries is effective in cases where NEs are included in the dictionary, it is often problematic when they are not included in them because new NEs are generated day by day.

We consider our bilingual NER to be a modified version of Hermjakob et al. 's method (2008) for adapting the monolingual NER. Our complete SMT systems are shown in Fig. 1.

1. Constructing transliteration models, we extract over 1,000 Japanese to English sub-string

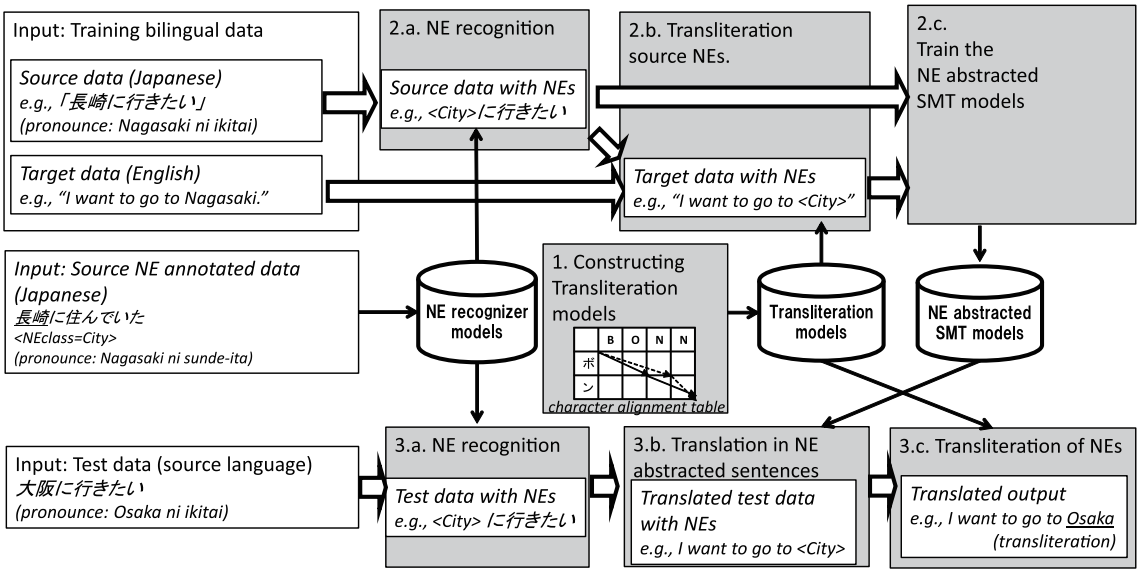

Figure 1: The flow of our system.

transliteration rules using a bilingual dictionary-based approach method (Saito et al., 2002). More concretely, for extracting these transliteration rules, the method utilized a character alignment table in accordance with entry pairs in the bilingual dictionary. Each sub-string alignment is generated by initial transliteration rules that have been relaxed for small differences about characters and each sub-string alignment is weighted cost as edit distance from transliteration rules to surfaces of words. The best total path in the character alignment table is then found, and if any edges included in the best path are not included in the rules, the path is added to the transliteration rules.

2. Training steps
   (a) Japanese NER extracts NE chunks from Japanese training corpus.
   (b) Our system uses transliteration models to align each Japanese NE chunk with the most phonetically similar English chunk whose similarity score exceeds a certain threshold. If no English chunk is aligned to a Japanese NE chunk, the label of that Japanese NE is removed in order to synchronize the number of NEs between the two languages.
   (c) Train the translation models using parallel corpus abstracted by NEs.
3. Test steps
   (a) Japanese NER extracts NE from Japanese test sentences.
   (b) The Japanese test sentences abstracted by NEs are translated into English sentences abstracted by NEs.
   (c) Transliterate the NEs into the target language. In the experiments in this work, we assume an oracle transliteration whereby we use English reference sentences to find the most similar chunks by the 2-b method because we do not focus on the transliteration itself. This part will be replaced by existing transliteration methods.

In the training steps, the NE alignment (2-b) is not developed in a straightforward way because there are NE suffix problems. For example, NER often recognizes "福岡 県" (pronounced "Fukuoka ken" and meaning "*Fukuoka prefecture*") as one NE "<Province>", although the suffix "県" (pronounced "ken" and meaning "*prefecture*") is not able to be transliterated. From this perspective, we move the NE suffix to outside the range of NE, such as "<Province> 県" and "<Province> prefecture", by hand-made rules. The number of these suffix rules can be reduced when we focus on fine-grain classes, unlike coarse-grain classes including multiple NE types.

To confirm the feasibility of this approach, we examined the performances of NERs (shown in Tables 2 and 3) for 100 sentences of Person and Location in Hiragana TIMES [1] (details in section 3.1). The original Japanese NER without restriction achieved the highest accuracy: 85.5 and 79.5 in F-value. Although the accuracy is degraded when NEs are restricted to having

[1] http://www.hiraganatimes.com/

Target language	NER method	Rec.	Prec.	F-value
Japanese	JPN-NER	0.762	0.975	0.855
Japanese	JPN-NER restricted by transliterated ENG.	0.576	0.926	0.710
English	JPN-NER and transliterated	0.609	0.968	0.748

Table 2: NER evaluation of "Person" in Hiragana TIMES.

Target language	NER method	Rec.	Prec.	F-value
Japanese	JPN-NER	0.660	1.000	0.795
Japanese	JPN-NER restricted by transliterated ENG.	0.639	1.000	0.780
English	JPN-NER and transliterated	0.639	0.930	0.757

Table 3: NER evaluation of "Location" in Hiragana TIMES.

an alignment to English, especially in the recall, the precision is still high in both languages, which makes it effective for NE abstracted SMT.

3. Experiments and Results

3.1. Experimental settings

We examined the effectiveness of NE abstraction methods with fine-grained NE labels compared to a non-NE abstraction method. Hiragana TIMES and BTEC (Kikui et al., 2006) were used as bilingual corpora. The details of data are shown in Table 4. Test set sentences (198 and 500 sentences each) including target NEs were randomly selected from the population of the test set (6,698 and 46,685 sentences each) because sentences rarely include target NEs, and the difference is hard to confirm when we compare with all sentences by statistical scores. These sentences were annotated with NE labels for analyzing the best performance when the NEs in the test set are completely extracted. GIZA++ (Och and Ney, 2003) was used for alignment words, Moses [2] was utilized as a phrase-based translation decoder, and minimum error rate training (MERT) was executed three times independently for tuning the models. BLEU and RIBES (Isozaki et al., 2010) scores calculated as arithmetic averages of three scores were used for the evaluation. The Japanese NER was trained by an enhanced version of Hashimoto's corpus (Hashimoto and Nakamura, 2010) with CRF by minimum classification error rate training (Suzuki et al., 2006).

[2]http://www.statmt.org/moses/

		Hiragana TIMES (news articles)		BTEC (travel dialogues)	
		Jpn.	Eng.	Jpn.	Eng.
Train	Sentence	172,740		397,565	
	avg-words	24.92	20.83	9.65	8.64
Develop	sentence	1,000		1,000	
	avg-words	23.66	21.34	9.30	8.31
Testset	sentence	198		500	
	avg-words	27.27	24.45	11.91	10.85

Table 4: Statistics of parallel corpora. The test set sentences of Hiragana TIMES include 100 sentences each for "Person" and "Location", with two sentences overlapping for a total of 198 sentences. Every test set sentence of BTEC includes "Location".

Method	Hiragana TIMES		BTEC	
	BLEU	RIBES	BLEU	RIBES
baseline (no-NER)	8.34	0.578	23.74	0.744
auto-NER (fine)	10.01	0.610	23.18	0.735
auto-NER (merge)	9.97	0.606	23.71	0.744
man-NER (fine)	10.86	0.616	25.33	0.753
man-NER (merge)	10.94	0.620	25.28	0.752

Table 5: Comparison of baseline and proposed abstraction with NEs by translation score.

3.2. Results and Analysis

In this section, we analyze the effectiveness of NE abstraction. The SMT scores of the automatic NE abstraction methods (with "auto-NER" as a header) and non-abstraction method (baseline) are shown in Table 5. In Hiragana TIMES, NE abstraction methods were effective. In contrast, in BTEC, NE abstraction methods were not effective because most of the sentences are typical dialogues that are easy to translate with the baseline, e.g., "*I must arrive in Tokyo by tomorrow morning* ." On the other hand, manual NER (indicated as "man-NER" in Table 5) achieved a better result than automatic NER methods ("auto-NER") and the baseline method, even in the BTEC corpus. These results indicate a potential improvement value of our method. A comparison between fine granularity ("auto-NER (fine)") and merged granularity ("auto-NER (merge)") shows that "auto-NER (merge)" achieved a higher score in BTEC and a competitive score in Hiragana TIMES. In manual NER, "man-NER (fine)" and "man-NER (merge)" results were almost the same. From these results, we conclude that "merge" is the best abstraction method in the current system.

Table 6 shows some examples translated by each of

Type	Example 1	Example 2
src. *Japanese* (pronounce)	子規 と 漱石 は 俳句 を 語り合 っ た 仲 。 (shiki to souseki wa haiku o kataria-tta naka)	1945 年 8 月 9 日 長崎 市 に 原子爆弾 が 投下さ れ ま した 。 (1945 nen 8-gatsu kokonoka nagasaki shi ni genshi-bakudan ga touka-sa re mashita)
src-NER	<Person> と <Person> は 俳句 を 語り合 っ た 仲 。	1945 年 8 月 9 日 <City> 市 に 原子爆弾 が 投下さ れ ま した 。
tgt-ref. *English*	shiki and soseki were friends who discussed haiku together .	on august 9 , 1945 , an atomic bomb was dropped on the city of nagasaki .
tgt-SMT *baseline*	子規 and the museum is haiku away with .	in 1945 on august 9 and nagasaki city in the atomic bomb was dropped .
tgt-SMT *withNER*	shiki and soseki is haiku away with .	in 1945 on august 9 , in nagasaki city , the atomic bomb was dropped .

Table 6: Improved examples of translation results by baseline SMT and NE abstraction SMT for Hiragana TIMES. The underlined words indicate NEs in the source language (Japanese) that have been translated into other underlined words in the target language (English).

Type	Example
src. *Japanese* (pronounce)	網走 で は 船 に 乗 っ て 約 1 時間 、 流氷 見学 ツアー を 楽し み ま した 。 (abashiri dewa fune ni notte yaku ichi-jikan , ryuuhyo tsua wo tanoshi mi mashita .
src-NER	網走 で は 船 に 乗 っ て 約 1 時間 、 流氷 見学 ツアー を 楽し み ま した 。 (no detected)
tgt-ref. *English*	*in abashiri , i got on a ship to join a one hour ryuuhyou-seeing tour .*
tgt-SMT *baseline*	網走 by boat in about 1 hour , 流氷 tours to enjoy .
tgt-SMT *withNER*	“網走 by boat in about 1 hour and 流氷 tours to enjoy .”

Table 7: A NOT improved example of translation results by baseline SMT and NE abstraction SMT for Hiragana TIMES.

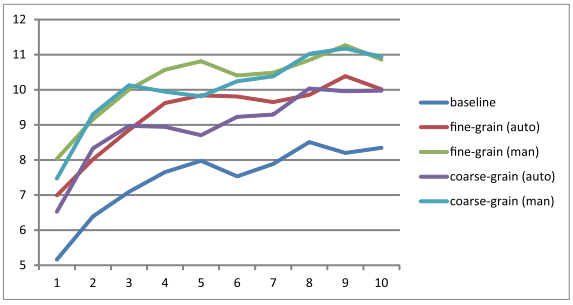

Figure 2: Learning curve of Hiragana Times (vertical: BLEU score, horizontal: amount of training data (x/10)).

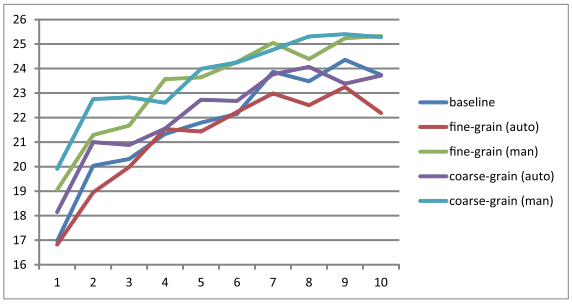

Figure 3: Learning curve of BTEC (vertical: BLEU score, horizontal: amount of training data (x/10)).

the SMT systems. In example 1, the baseline could not translate “子規” because it was processed as an unknown word. In example 2, “長崎” and “*nagasaki*” were correctly translated, but the context was incorrect. The proposed method handled these two examples better than the baseline. However, there is an example where the proposed method did not perform well. The example in Table 7 is a case of NE not being translated due to a miss-detect in recognizing “網走 (abashiri)” as an NE. Improvements to NER should resolve this type of problem.

Learning curves are shown in Figs. 2 and 3. With a small amount of training data (10% -80%), NE abstraction methods exceeded the baseline method even in the BTEC corpus. This confirms that when the amount of the parallel corpus is small for training, NE abstraction methods are effective.

3.3. Additional approaches for BTEC corpus

For improving the SMT performance using a full amount of BTEC, we introduced two additional approaches. In the first approach, we added 3 million parallel corpora belonging out of domain, the results of which are shown in Table 8. The value of improve-

Training data	Method	BLEU
BTEC only	baseline(no-NER)	23.74
	auto-NER (merge)	23.71
	man-NER (merge)	25.28
BTEC+3M	baseline(no-NER)	24.34 (+0.60)
	auto-NER (merge)	24.64 (+0.93)
	man-NER (merge)	25.48 (+0.20)

Table 8: Comparison of in-domain corpus and adding huge out-domain corpus translation score.

Method	All sentences	High freq.	Low freq.
baseline	28.39	29.09	20.19
auto-NER	24.59 (-3.80)	22.25 (-6.84)	21.40 (+1.21)
man-NER	26.67 (-1.72)	20.42 (-8.67)	23.12 (+2.93)
combination (auto-NER)	28.70 (+0.31)	29.09	21.40
combination (man-NER)	29.14 (+0.75)	29.09	23.12

Table 9: The result of NE abstraction depends on the term frequency in training data. "All sentences" is limited to 100 sentences. "High freq. " and "Low freq. " are selected from among 29 sentences of "All sentences".

ment in the NER abstraction ($+0.93$) was larger than that in the baseline ($+0.60$), although the difference is not significant.

In the second approach, we utilized the threshold of term frequency of each NE (Hermjakob et al., 2008), dividing all sentences into two sets including one high and one low frequency NE. Table 9 shows the result when we set the threshold to 100. NE abstraction methods are effective in low frequency data (auto:$+1.21$, man:$+2.93$), while in contrast, they perform poorly in high frequency data (auto:-6.84, man:-8.67). When NE abstraction is automatically limited for low frequency NE ("combination"), we can achieve better results than the baseline method (auto:$+0.31$, man:$+0.75$).

4. Conclusion

We proposed and analyzed a name translation method for alleviating data sparseness of NEs in SMT. Our method is based on a bilingual named entity recognizer depending on monolingual fine-grained NER and transliteration. We reduced the degrading effect of NER mistakes by focusing on a few fine-grained labels that are frequently used and easy to recognize classes in target corpus. Experimental results demonstrate that NE abstraction methods are effective in cases where data are sparse.

In future work, we will compare the proposed method with the coarse-grain NE and analyze the out of scope NE target that is not derived from transliteration.

5. Bibliographical References

Al-Onaizan, Y. and Knight, K. (2002). Machine Transliteration of Names in Arabic Text. In *Proceedings of the Association for Computational Linguistics Workshop on Computational Approaches to Semitic Languages*.

Hashimoto, T. and Nakamura, S.-i. (2010). Kakucho-Koyuu-Hyougen tag Tsuki Corpus no kouchiku - Hakusho, Shoseki, Yahoo! Chie-bukuro core data (in Japanese). In *Proceedings of the 16th Annual Meeting of the Association for Language Processing*, pages 916–919.

Hassan, A., Fahmy, H., and Hassan, H. (2007). Improving named entity translation by exploiting comparable and parallel corpora. In *RANLP, AMML Workshop*.

Hermjakob, U., Knight, K., and Daumé III, H. (2008). Name Translation in Statistical Machine Translation earning When to Transliterate. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 389–397.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Kikui, G., Yamamoto, S., Takezawa, T., and Sumita, E. (2006). Comparative Study on Corpora for Speech Translation. *IEEE Transactions on audio, speech, and language processing*, 14(5):1674–1682.

Knight, K. and Graehl, J. (1997). Machine Transliteration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 128–135.

Li, H., Zheng, J., Ji, H., Li, Q., and Wang, W. (2013). Name-aware Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 604–614.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Och, F. J. (1999). An efficient method for determining bilingual word classes. In *Proceedings of the*

ninth conference on European chapter of the Association for Computational Linguistics, pages 71–76.

Saito, K., Shinohara, A., Nagata, M., and Ohara, H. (2002). A Transliteration Algorithm for Adapting a Japanese Voice Controlled Browser to English (in Japanese). *Transactions of the Japanese Society for Artificial Intelligence*, 17(3):343–347.

Sekine, S. (2008). Extended named entity ontology with attribute information. In *Proceedings of the 6th International Language Resources and Evaluation*.

Suzuki, J., McDermott, E., and Isozaki, H. (2006). Training Conditional Random Fields with Multivariate Evaluation Measures. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 217–224.