Translation of New Named Entities from English to Chinese

Zizheng Zhang and **Tosho Hirasawa** and **HouJing Wei** and
Masahiro Kaneko and **Mamoru Komachi**
Tokyo Metropolitan University
6-6 Asahigaoka, Hino
Tokyo 191-0065, Japan
{zizheng,houjing}@komachi.live,
{hirasawa-tosho,kaneko-masahiro}@ed.tmu.ac.jp, komachi@tmu.ac.jp

Abstract

New things are being created and new words are constantly being added to languages worldwide. However, it is not practical to translate them all manually into a new foreign language. When translating from an alphabetic language such as English to Chinese, appropriate Chinese characters must be assigned, which is particularly costly compared to other language pairs. Therefore, we propose a task of generating and evaluating new translations from English to Chinese focusing on named entities. We defined three criteria for human evaluation—fluency, adequacy of pronunciation, and adequacy of meaning—and constructed evaluation data based on these definitions. In addition, we built a baseline system and analyzed the output of the system.

1 Introduction

A machine translation (MT) system is expected to generate the correct translation results for each input. However, new named entities (NEs), such as company names, character names, and product names, are constantly being created worldwide. Therefore, such words must be assigned new translations without referring to any translations in other languages.

In particular, the translation of NEs between different alphabets, for example, from English (En) to Chinese (Zh) characters, is more difficult than that between other language pairs. It is necessary to select appropriate Chinese characters (Hanzi) in consideration of appropriate fluency, adequacy of pronunciation (hereinafter referred to as "pronunciation"), and adequacy of meaning (hereinafter referred to as "meaning"). These three dimensions should also be considered in NE translation evaluation. For example, the NE pair of (Curtiss-Wright, 柯蒂斯-莱特) is evaluated high in terms of pronunciation. However, its fluency in Chinese is not good because it is not an original Chinese word.

Although several studies have been conducted on En-Zh MT (Wang et al., 2017; Deng et al., 2018), NE translation (Chen and Zong, 2011) and transliteration (Wan and Verspoor, 1998; Benites et al., 2020), no research has been conducted so far on generating and evaluating the translations of brand new NEs in terms of fluency, pronunciation, and meaning. The difficulty in considering these three dimensions makes translating a new NE a challenging task. In Chinese, there can be several Hanzi with similar pronunciations or meanings, and they all can be selected for appropriate NE translation. For instance, the NE pairs in En-Zh of (Blackstone Group, 黑石集团) and (Blackstone Group, 百仕通), where the former represents the literal translation and the latter represents transliteration, are both correct, and it is difficult to judge which is preferable. Thus, it is first necessary to define the criteria of fluency, pronunciation, and meaning.

Thus, in this paper, we propose a novel task of generating and evaluating brand new NE translations for En-Zh. The main contributions are:

- We propose the evaluation criteria for new En-Zh NE (company name) translations—fluency, pronunciation, and meaning.
- We create a baseline model for NE translations and analyze the results.
- We provide and release a novel method of evaluation dataset [1] for En-Zh, focusing on company names, which includes both real NE translations and our system output.

2 Related Work

In terms of NE translation (Chen et al., 1998; Wan and Verspoor, 1998; Oh et al., 2009), because the two languages use completely different symbolic representations in terms of graphemes and

[1] https://github.com/toshohirasawa/enzh-named-entity-translation

Proceedings of the 7th Workshop on Asian Translation, pages 58–63

Score	Fluency	Pronunciation	Meaning
5	Original **AND** #splitting= 0	Similar **AND** #syllables are close	Translated **AND** Shortly
4	Original **AND** #splitting= 1	-	-
3	Original **AND** #splitting= 2	Similar	Translated
2	Original **AND** #splitting≥ 3	-	-
1	Others	Others	Others

Table 1: Criteria for human evaluation are fluency, pronunciation, and meaning. With respect to fluency dimension, **Original** indicates that there is at least one original Chinese word/phrase in Chinese NE, and **#splitting** refers to number of semantic splits. With respect to pronunciation dimension, **Similar** indicates that pronunciation in En-Zh is similar, and **#syllable** represents number of syllables in En-Zh. With respect to meaning dimension, **Translated** denotes that all words are translated directly, and **short** denotes that number of Hanzi is equal to or less than 4.

English	Chinese (Pinyin)	F	P	M
Celanese	塞拉尼斯 (Sai La Ni Si)	1	5	1
Altria	奥驰亚 (Ao Chi Ya)	3	5	1
Apple Inc.	苹果公司 (Pin Guo Gong Si)	5	1	5

Table 2: Examples of evaluation data, in which two annotators evaluated identically. **F**, **P**, and **M** denote fluency, pronunciation, and meaning, respectively.

phonemes, the English graphemes, phonetically associated English letters, must be converted into Hanzi, which represent ideas and meanings. In terms of literal translation, because Hanzi usually express certain connotations, choosing the appropriate Hanzi should also be considered. In addition, owing to the lack of apparent semantic content on location names and people's names, these words cannot be expressed in Chinese through words equivalent in meaning. Further, it is very likely that the standard translation of these words cannot be found in existing lexical resources, which increases the complexity of the task.

A semantic transliteration method (Li et al., 2007) is proposed for the translation of personal names from English to Chinese, which considers the language of origin, gender, and the given or surname information of the source names. The approach aims at maintaining the phonetic equivalence as well as optimizing the semantic transfer.

However, as highlighted in the paper, the research is a case study, and the proposed mathematical framework does not extend to the machine transliteration of NEs.

3 Dataset

We construct our evaluation data based on the company list of the New York Stock Exchange [2], in which we select the companies that have both English and Chinese Wikipedia pages as the source and target NEs, respectively; the titles of Chinese pages can be requested for by using the Langlinks API and the English titles. We chose company names because they reflect the corporations' characteristics, providing more information for evaluating fluency, pronunciation, and meaning. Because we focus on the English to Chinese NE translation, companies from Greater China are ignored. In all, 338 En-Zh NE pairs were evaluated by two annotators in terms of the three dimensions mentioned above.

3.1 Annotation

Tables 1 and 2 list the criteria used for human evaluation as well as some examples of our evaluation data. As a global criterion, we ignore certain common words that do not contribute to the translation of business names, such as Inc., corporation, and group.

When evaluating the performance of MT, different types of human judgment including fluency and adequacy are employed. A quantitative and qualitative investigation was conducted by (Tu et al., 2017). They confirmed that the source and target contexts in neural MT are highly correlated with translation adequacy and fluency, respectively. For our study, this finding may indicate that the more common the translation using existing Chinese expressions, the better is its fluency. As for

[2] https://en.wikipedia.org/wiki/Category:Companies_listed_on_the_New_York_Stock_Exchange

adequacy, the consistency between the source and target contexts should be prioritized in terms of both pronunciation and meaning.

Fluency measures whether a translation is fluent, regardless of the correct meaning, but it takes the order in the translation highly into account (Snover et al., 2009). We use a 5-level scale to evaluate the fluency in Chinese, where two dimensions are considered. Considering that the original words or phrases in the target language provide more fluency, we make one dimension as to whether the original Chinese words or phrases are included.

Moreover, less semantic splitting provides greater fluency because the more similar the modification relationship among Hanzi, the more likely they are not to be split. Another dimension we consider is the number of semantic splits. In addition, a missing is considered for the semantic orientation of subtokens (the result of semantic splitting). If there is at least one combination between subtokens consisting of (positive word, negative word) or (neutral word, negative word), the fluency score is decreased by 1 to obtain the final fluency score (which should be at least 1). For example, "罗渣士通讯" is one way to translate "Rogers Communications", where it will be split three times to give "罗 | 渣 | 士 | 通讯" , which implies that it will obtain a fluency score of 2. As the subtoken "渣" is a negative word, whereas others are neutral, the missing leads to a score of -1 so that the final score for "罗渣士通讯" is 1.

Adequacy measures whether the translation conveys the correct meaning, even if the translation is not completely fluent (Snover et al., 2009). In this study, we use a three-level scale to measure the translation performance with respect to both pronunciation and meaning because there is no necessity to subdivide further. For the meaning dimension, we consider it being short as a criterion because short meanings are easy to remember, which is essential for a business name. It should be noted that the names of people, places, and so on in the transliteration (pronunciation) should also be evaluated with a high score in the translation (meaning) because they also contribute to the meaning.

3.2 Agreement

We evaluated the annotations across two annotators using the kappa coefficient (Landis and Koch, 1977). The kappa coefficients of fluency, pronunciation, and meaning are approximately 0.68, 0.62, and 0.65, respectively, which indicates that the inter-rater reliability is substantial.

4 Baseline Model

In the following section, we describe a baseline model for character-based NE translation. Furthermore, we propose filters to remove noisy samples from the Wikititles dataset and demonstrate that the sanitized data could improve the performance of the model.

4.1 Model

The attention-based encoder–decoder model is a well-known architecture for MT (Bahdanau et al., 2015; Vaswani et al., 2017). The model tackles MT as a sequence-to-sequence problem.

Although the model was first proposed to operate at the level of words, recent papers have proposed character-level neural MT models (Lee et al., 2017). In the present study, we employed Bahdanau et al. (2015) as our baseline model for English–Chinese NE translation.

4.2 Experimental Setup

Model The encoder of our model has two layers with 256 hidden dimensions; therefore, the bidirectional GRU has a dimension of 512 and the decoder GRU state has a dimension of 256. The input word embedding and output vector sizes are 256.

For training, we used the Adam (Kingma and Ba, 2014) optimizer with a learning rate of 0.001, clipping gradient norm of 1.0, dropout rate of 0.5, batch size of 512, and early stopping patience of 10. In the evaluation phase, we performed a beam search with a size of 12. We trained three models with different seeds and used character-level BLEU to evaluate the model.

Dataset We train and validate our models on a subset of the Wikititles dataset from which parallel entities are removed, and evaluate them on the Wikititles company dataset. In particular, we randomly split the Wikititles into two parts: 99% for training and 1% for validation. It should be noted that the training and validation sets include NEs from various domains, whereas the test set includes only company names.

Further, we trained and evaluated our models on sanitized data, that is, data from which the samples satisfying any of the following four conditions were removed: 1) English and Chinese names are identical, 2) Chinese name does not contain Hanzi,

Split	Examples	Char (En)	Char (Zh)
Train	827,604	38.8	13.2
Validation	8,357	38.9	13.2
Sanitized Data			
Train	690,613	37.2	10.9
Validation	6,975	37.5	11.0
Test	338	32.7	10.7

Table 3: Statistics of dataset used to train MT models. **Char (En)** and **Char (Zh)** denote the average number of characters in the English and Chinese words for the entities, respectively.

Model	Validation BLEU	Test BLEU	chrF
Vanilla	48.59	8.07±0.36	15.84
Sanitized	49.03	18.37±1.82	21.16

Table 4: Named entity translation performance of baseline models. Character-level BLEU and chrF scores are reported. "Vanilla"/"Sanitized" denotes model trained on original/sanitized training data.

English	System	B	F	P	M
Zoetis	佐伊蒂斯	0.00	1	5	1
Wipro	维普罗	50.81	1	5	5

Table 5: Translation examples generated by the "sanitized" model. **B**, **F**, **P**, and **M** denote BLEU, fluency, pronunciation, and meaning, respectively.

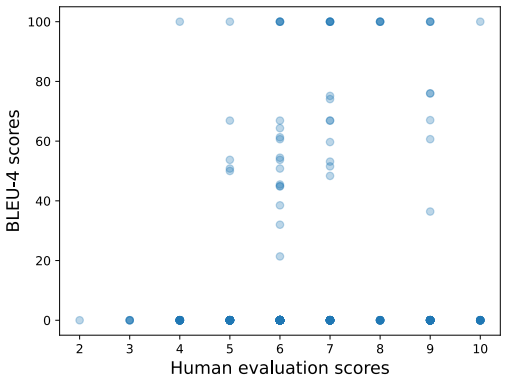

Figure 1: Correlation between BLEU score and human evaluation scores. Darkness of point denotes the level of overlap.

3) English name contains Hanzi, and 4) English or Chinese name is longer than 50 or 20 characters, respectively.

Table 3 presents the statistics of the resulting training and validation data. All entities in both English and Chinese are split into characters, and the space is replaced with a special token (in our case, <s>). The vocabularies are built with all words from the original/sanitized training data, yielding 1,063/353 characters in English and 9,598/8,852 in Chinese.

4.3 Results and Analysis

General performance Table 4 presents the corpus-level BLEU-4 and chrF-4 scores (Popović, 2015) for each model on the English to Chinese translation. We trained three models with random initial states and used their average BLEU scores with error range to represent the final BLEU score on the test set. It can be observed that our baseline system achieved a reasonable performance (∼50 BLEU) on the validation set but failed to translate most entities in the test set (<20 BLEU).

The poor performance of our model is attributed to the fact that the Wikititles dataset contains highly diverse data, whereas the test set includes only company names. Therefore, it is difficult to train an NE translation model and generate new company names.

Another potential reason is that the Wikititles dataset contains noisy data. The results obtained for models trained on sanitized data ("Sanitized" in Table 4) support these ideas and reflect a substantial improvement (+10.30 BLEU) obtained using four simple rule-based filters.

Human evaluation Further, we manually annotated 338 outputs of the model trained with sanitized data using the criteria introduced in Sec. 3.1.

Figure 1 depicts the correlation between the BLEU score and human evaluation scores, where we use the sentence-level BLEU-4 score (Papineni et al., 2002) as the BLEU score for each translation item. We set $F + \max(P, M)$ as the human evaluation scores, where F, P, and M represent fluency, pronunciation, and meaning, respectively. Here, considering that certain NEs are translated using hybrid methods, $\max(P, M)$ is used to balance the weights from transliteration and literal translation. The Pearson correlation coefficient is also calculated, as approximately 0.12. Both Figure 1 and the Pearson correlation coefficient indicate that there is nearly no correlation between these two scores.

The reason for the low correlation is that most of the translation obtains a BLEU score of 0.0 but different human evaluation scores. In the present test set, certain NE translations are not similar to the references but could be evaluated by humans as being effective. Similarity with references does not represent the quality of the NE translation.

Table 5 presents translation examples generated by the model trained on the sanitized data. The reference NE of "Zoetis" is "硕腾", where our model transliterates it and makes a high P score. It obtains a low F score because there is no original Zh word. For the NE "Wipro", our model translated it with a homophone, where the reference NE is "威普罗"; as a transliteration of people's names, it obtained high P and M scores.

5 Conclusion

This paper describes a new method for NE translation from English to Chinese. For this purpose, we presented human evaluation criteria for business names and build a test set. Further, we found that the correlation between the BLEU score and human evaluation scores is weak. The reason is that while the human evaluation scores represent the quality of the NE translation, BLEU represents the similarity between the reference NEs and the outputs from the model. Thus, we conclude that, for evaluating NE translations, reference-less methods should be more effective than reference-based methods.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020. TRANSLIT: A large-scale name transliteration resource. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France. European Language Resources Association.

Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Chung Tsai. 1998. Proper name translation in cross-language information retrieval. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 232–236, Montreal, Quebec, Canada. Association for Computational Linguistics.

Yufeng Chen and Chengqing Zong. 2011. A semantic-specific model for Chinese named entity translation. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 138–146, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, Changfeng Zhu, and Boxing Chen. 2018. Alibaba's neural machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 368–376, Belgium, Brussels. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Haizhou Li, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong. 2007. Semantic transliteration of personal names. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 120–127, Prague, Czech Republic. Association for Computational Linguistics.

Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Can Chinese phonemes improve machine transliteration?: A comparative study of English-to-Chinese transliteration models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 658–667, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Stephen Wan and Cornelia Maria Verspoor. 1998. Automatic English-Chinese name transliteration for development of multilingual resources. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1352–1356, Montreal, Quebec, Canada. Association for Computational Linguistics.

Yining Wang, Long Zhou, Jiajun Zhang, and Chengqing Zong. 2017. Word, subword or character? An empirical study of granularity in Chinese-English NMT. In *China Workshop on Machine Translation*, pages 30–42. Springer.