Improved Named Entity Translation and Bilingual Named Entity Extraction

Fei Huang, Stephan Vogel
Interactive System Labs, Carnegie Mellon University
Pittsburgh, PA 15217
{fhuang, vogel, ahw}@cs.cmu.edu

Abstract

Translation of named entities (NE), including proper names, temporal and numerical expressions, is very important in multilingual natural language processing, like crosslingual information retrieval and statistical machine translation. In this paper we present an integrated approach to extract a named entity translation dictionary from a bilingual corpus while at the same time improving the named entity annotation quality.

Starting from a bilingual corpus where the named entities are extracted independently for each language, a statistical alignment model is used to align the named entities. An iterative process is applied to extract named entity pairs with higher alignment probability. This leads to a smaller but cleaner named entity translation dictionary and also to a significant improvement of the monolingual named entity annotation quality for both languages. Experimental result shows that the dictionary size is reduced by 51.8% and the annotation quality is improved from70.03 to 78.15 for Chinese and 73.38 to 81.46 in terms of F-score.

1. Introduction

Translation of named entities (NE), including proper names, temporal and numerical expressions, is very important in statistical machine translation (SMT), because named entities, especially named persons, locations and organizations, convey essential meaning in human languages [1][2]. Some approaches, like word/subword translation or transliteration, have been explored in the past few years [3][4][5]. However, applying the word-based source-channel paradigm to named entities translation usually leads to unsatisfactory results. The reason is, while the translation is conducted on word or character level (e.g., translation from Chinese to English), the meaning of a single word or character is inappropriately considered and some inherent properties of named entities are disregarded. For example, when translating "风陵渡" (a Chinese location name) to English, the correct translation should be "Fenglingdu", but the character-by-character translation is "wind tomb cross", which makes no sense in the given context. Template-based NE translation may work well for temporal and numerical NEs, because of their limited vocabulary and fixed usage, but does not generalize well enough for proper name translation, especially foreign location or person names.

One possible solution is to build a bilingual named entity dictionary. Whenever a named entity is detected in the source language, its corresponding translations in the target language are acquired by dictionary lookup, and plugged into the appropriate position in the translation output. To build such a named entity dictionary, this approach needs a sentence aligned bilingual corpus with named entity annotation. Given the corpus, the dictionary can be built through named entity alignment. However, it is not easy to obtain such an annotated corpus. Manual annotation of bilingual corpora is extremely expensive, and automatic annotation using commercial software cannot guarantee high quality in named entity annotation, although it can be good enough for the starting point of an iterative procedure.

In this paper we propose an iterative approach to named entity translation/named entity extraction to a bilingual Chinese/English corpus. The initial bilingual corpus is first annotated using commercial NE annotation software, whose output is the baseline annotation corpus. Then an alignment model is applied to this corpus to generate a baseline NE dictionary. After that, the dictionary is used to correct some annotation errors in the corpus, and a new dictionary is generated from the corrected corpus. This procedure is iteratively conducted until there is no further improvement in the dictionary and annotation quality.

The structure of this paper is as follows: in section 2 the NE alignment model will be discussed, in section 3, the corrective annotation model will be proposed. Section 4 presents the whole iterative procedure, and discusses the experiment setting and results. Conclusions will be given in the last section.

2. Named Entity Alignment Model

The NE alignment model is exploited to generate a bilingual NE dictionary. For each NE entry in the source language, the dictionary contains $\boldsymbol{m}$ most probable NE translations in the target language. These candidate translations are obtained according to the co-occurrence

frequency among aligned NE pairs with minimum alignment cost in a sentence.

2.1. Named entity translation cost

Given a sentence aligned bilingual text, word translation probabilities $p(f \mid e)$ can be estimated using the well-known alignment models [6][7]. Such a probability distribution can then be used to calculate the probability that a Chinese NE is the translation of an English NE.

Let NE_e denote an English named entity, which is composed of I English words, $e_1, e_2, ...e_I$, and let NE_c denote a Chinese named entity, which is composed of J Chinese words, $c_1, c_2, ...c_J$. The translation probability of the named entities pair $P(NE_c \mid NE_e)$ is computed using the IBM model-1, as:

$$P_{trans}(NE_c \mid NE_e) = \frac{1}{I^J} \prod_{j=1}^{J} \sum_{i=1}^{I} p(c_j \mid e_i) \quad (1)$$

This alignment model is asymmetric, as one source word can be aligned to one target word only, while one target word can be aligned to more than one source words. Therefore, we estimate both $P(NE_c \mid NE_e)$ and $P(NE_e \mid NE_c)$, and define the NE translation cost as:

$$\begin{aligned} & C_{trans}(NE_e, NE_c) \\ & \equiv C_{trans}(NE_e \mid NE_c) + C_{trans}(NE_c \mid NE_e) \\ & \equiv -[\log P_{trans}(NE_e \mid NE_c) + \log P_{trans}(NE_c \mid NE_e)] \end{aligned} \quad (2)$$

That is, the translation cost of a given NE pair (NE_e, NE_c) is composed of translation cost from NE_e to NE_c, and the cost of the reverse translation.

2.2. Sentence level named entity alignment

The sentence level NE alignment is to find a NE alignment scheme for a given bilingual sentence pair, to minimize the sentence alignment cost, SAC, which is defined as the sum of the translation cost of those aligned NE pairs.

Mathematically, let $E = (NE_{e1}, NE_{e2}, ...NE_{em})$ denote the set of m NEs in the given English sentence, and $C = (NE_{c1}, NE_{c2}, ... NE_{cn})$ denote the set of n NEs in the given Chinese sentence. The optimal NE alignment scheme A_{opt} satisfies

$$\begin{aligned} A_{opt} &= \arg\min_A SAC(A) \\ &= \arg\min_A \sum_{\substack{NE_e \in E, NE_c \in C \\ (NE_e, NE_c) \in A}} C_{trans}(NE_e, NE_c) \end{aligned} \quad (3)$$

To find A_{opt}, an algorithm similar to the competitive linking algorithm [8] is adopted:

1. Initialize *NE-Aligned* to be an empty set and *NE-Pairs a*s the list of all possible combinations ($m \times n$ entries) of a source language NE and a target language NE in the given sentence pair;
2. Sort *NE-Pairs* in ascending order according to their translation cost;
3. Move the topmost pair *(NE_e, NE_c)*, i.e. the pair with the smallest translation cost $C_{trans}(NE_e, NE_c)$, from *NE-Pairs* to *NE-Aligned*;
4. Remove all *(NE_e, •)* and *(•, NE_c)* from *NE-Pairs*;
5. Repeat from Step 3 until *NE-Pairs* is empty. The resultant *NE-Aligned* leads to the A_{opt}.

Note that this algorithm is a greedy approximation, so it cannot guarantee the global optimality of the alignment. But empirically it often finds the alignment with minimum or close to minimum sentence alignment cost.

2.3. Corpus level named entity alignment probability

The sentence level NE alignment is conducted over the whole bilingual corpus. For each source language named entity, all the aligned named entities in the target language (over the whole corpus) are stored, together with the frequencies of their alignment.

The NE alignment probability is then just the normalized alignment frequencies:

$$P_{align}(NE_e \mid NE_c) = \frac{freq(NE_e, NE_c)}{\sum freq(\bullet, NE_c)} \quad (4)$$

Thus the entry in the dictionary is a triple $(NE_c, NE_e, P_{align}(NE_e \mid NE_c))$.

Since the alignment is bi-directional, formula (4) can also be used to estimate $P_{align}(NE_c \mid NE_e)$. The NE alignment cost is then symmetrically defined as:

$$\begin{aligned} & C_{align}(NE_e, NE_c) \\ & \equiv \log P_{align}(NE_e \mid NE_c) + \log P_{align}(NE_c \mid NE_e) \end{aligned} \quad (5)$$

3. Corrective Named Entity Annotation

Given the NE translation dictionary, some tagging errors in the baseline annotation can be corrected, by augmenting monolingual annotation with cross-lingual information. However, considering noisy errors in the NE translation dictionary, mismatches in sentences alignment, even the inexact translation among correctly aligned sentence pair, the annotation which is solely based on the NE dictionary will result in lower recall,

although higher precision. So the corrective approach will adopt the new annotation only when the sentence alignment cost is lower than the baseline's cost.

Now with the NE translation probability which expresses the context-independent alignment cost between a NE pair, and the NE alignment probability which indicates their alignment cost in the context of the whole bilingual corpus, the combined alignment cost, which we call the *augmented NE alignment cost* C_{aug} is defined as:

$$C_{aug}(NE_e, NE_c) = \lambda C_{trans}(NE_e, NE_c) + (1-\lambda) C_{align}(NE_e, NE_c), \quad (6)$$

where C_{trans} is defined as in formula (2), and C_{align} is defined as(5). The interpolation parameter λ is selected to be 0.5 in the current implementation.

NEs can be tagged with wrong TYPE tags, so the match between different TYPE NEs, e.g. a LOCATION NE is aligned to an ORGANIZATION NE, is allowed, but with a lower probability. Similar to the IBM-2 model, position information is also incorporated into the cost estimation, but with a small weight only because of the significant difference of word ordering between Chinese and English.

Since the NE translation probability is computed from word translation probability (see formula (1)), which in turn is computed from their co-occurrence frequency, the alignment cost between two longer NEs is always larger than that of two sub-NEs that are part of the longer NEs. For example, the shorter NE pair "香港" and "Hong Kong" co-occur more frequently than the longer pair, "香港 特别 行政区" and "Hong Kong Special Administrative Region", and whenever the latter NE pair co-occurs the former ones will also co-occur. In such a case, the longer NE-pair has no chance to be aligned because of the higher alignment cost. To deal with this problem, a "length bonus" is applied to the alignment cost computation. That is, the alignment cost is discounted proportionally to the length of aligned NE pairs.

Then the overall sentence alignment cost is computed as in formula (3), but C_{trans} is replaced with C_{aug}, resulting the *augmented sentence alignment cost* (*ASAC*).

Therefore, the corrective NE annotation scheme is:

1. Compute *ASAC* on the baseline annotation;
2. Tag the sentence pair with all possible annotations, that is, find all matching NEs from the baseline corpus;
3. Find the alignment with minimum *ASAC*, using the same greedy approximation algorithm as in 2.2. If this alignment cost is less than the baseline cost (computed in step 1), accept the alignment and the corresponding annotation; otherwise keep the original annotation.
4. For unaligned but frequent NEs, tag them with their most frequent TYPEs to reduce the side effect from inaccurate sentence alignment or inexact translations.

4. Iterative NE Alignment and Annotation Experiment

The bilingual corpus used in the experiment is the Hong Kong News Corpus, distributed through the Linguistic Data Consortium, which contains 96,320 sentence pairs, 3,034,253 English words, and 3,008,665 Chinese words. The Chinese sentences are pre-segmented using a maximum matching segmenter with a wordlist of 170K words. The segmentation slightly degrade the baseline annotation quality, but the reduction is quite limited, with only 1~2% in terms of F-score. Considering the necessity of building the translation lexicon and the improvement from the proposed iterative approach, such a reduction is acceptable.

The baseline bilingual annotation is achieved by BBN's named entity annotation software, IdentiFinder™[2]. The tagged named entities include 7 categories, person name, location name and organization name, date/time expression, and money/percentage expression. The last four categories are relatively easy and reliable to annotate with rule-based approach, because of their regularity (limited vocabulary, fixed usage). So we will focus on the first three categories, i.e., named person, location and organization.

Given the annotated bilingual corpus, the NE alignment procedure and corrective annotation procedure are iteratively applied, to construct the NE translation dictionary and improve the NE annotation in turn. After each iteration, the NE dictionary has less entries but a more accurate translation probability, and more and more errors in the annotated corpus are corrected.

To evaluate the annotation accuracy, a test set is randomly selected from the whole corpus, which contains 192 sentence pair, 12430 words. In these sentences 73 person names, 182 location names and 193 organization names were found and manually annotated according to the HUB-4 NE annotation guideline [9]. The automatically generated annotation was then evaluated by calculating precision and recall with respect to this gold standard. Precision is defined as

$$P = \frac{\#\,of\ correct\ annotated\ NEs}{\#\,of\ all\ annotated\ NEs}.$$

Recall is defined as

.

Table 1. Dictionary size and monolingual annotation accuracy after each iteration

Iteration	Dictionary Size	Chinese			English		
		Precision	Recall	F-score	Precision	Recall	F-score
baseline	N/A	75.00	65.68	**70.03**	76.05	70.89	**73.38**
1	41397	80.41	71.89	**75.91**	77.11	83.11	**80.00**
2	27874	79.26	75.56	**77.37**	77.58	84.01	**80.67**
3	27559	80.21	75.56	**77.82**	77.82	84.31	**80.94**
4	27324	79.17	76.84	**77.99**	78.27	84.92	**81.46**
5	27264	82.07	74.58	**78.15**	78.27	84.92	**81.46**

$$R = \frac{\#\,of\ correct\ \ annotated\ \ NEs}{\#\,of\ \ all\ \ correct\ \ NEs}$$

The F-score, a combined measure of NE annotation's precision and recall, is defined as

$$F = \frac{2PR}{P+R}.$$

Because some frequent tagged NEs in the baseline, like "政府/Government", are not in accordance with the NE definition used for evaluation, those incompatible NEs were removed from the baseline annotation.

Table 1 demonstrates the size of the NE dictionary and the monolingual annotation accuracy after each iteration. The baseline is given by the bilingual corpus where source and target sides are tagged independently. Using bilingual information, i.e., having source language and target language tagging influence each other through the alignment, gives a considerable improvement in precision and recall for both languages. This in turn leads to a cleaner lexicon which is much smaller, with only 65% entries of the first dictionary, as many entries with wrongly tagged NEs are removed. Further iterations give an additional small but still noticeable improvement.

Figure 1 presents some examples from the dictionary, with corresponding translations in the 1st and 5th iteration, where it can be found that after each iteration the translation probability mass gradually transfers to the correct NEs. Notice that one Chinese NE can have multiple English translations, e.g. "陈方安生" can be translated as "Anson Chan", "Mrs Chan" or just "Chan", all of which are correct translations depending on the given context. In these cases, the probability mass is distributed according to their co-occurrence frequency. Wrong translations such as "Patrick Lau" might be from mismatching annotations where they are the only tagged NEs to be matched.

Figure 2 illustrates one annotated sentence pair from the corpus, with the baseline annotation and the annotation after the 5th iteration. Three kinds of NE annotation errors can be found:

- Incorrect annotation: for example, the LOCATION "中华人民共和国 (People's Republic of China)" is tagged as "中华人民共和国 香港", which indeed includes part of the second named entity, "Hong Kong";
- Missing annotation: for example "HKSAR" is not tagged in the baseline;
- Spurious annotation: for example, "Administrative Region" is falsely tagged as an ORGANIZATION named entity.

The presented example shows that with the NE dictionary generated from the alignment model, some annotation errors are corrected, such as "People's Republic of China" which now is aligned to "中华人民共和国" rather than "中华人民共和国 香港", and "香港 会议 展览 中心" is aligned to "Hong Kong Convention and Exhibition Centre" rather than "Hong Kong Convention and Exhibition Centre (HKCEC".

However, there are a number of cases where the tagging of the baseline system is consistently wrong. For example, "Hong Kong Special Administrative Region" is always tagged as "Hong Kong" and "Administrative Region". These errors cannot be corrected by the iterative approach as the baseline NE dictionary gives a high probability for the wrong NE-to-NE alignment.

6. Conclusion

We presented an integrated approach to extract a named entity translation dictionary from a bilingual corpus while at the same time improving the named entity annotation quality. Starting from the bilingual corpus where the named entities were extracted independently for each language, a statistical alignment model was used to align the named entities. An iterative process was applied to extract named entity pairs with higher alignment probability. This resulted in a smaller but cleaner named entity translation dictionary and also in a significant improvement of the monolingual named entity annotation quality for both languages. Experimental result showed that the dictionary size was

Baseline NE Dictionary		NE Dictionary after the 5th iteration	
LOCATION: 澳大利亚			
Australia (√)	0.636	Australia (√)	0.867
Mutual Legal Assistance	0.182	TSE	0.066
Tim Fischer	0.045	John Olsen	0.033
Council	0.045	Tim Fischer	0.033
TSE	0.045		
Jeff Kennett	0.045		
ORGANIZATION: 创新 科技 委员会			
Commission on Innovation and Technology (√)	0.373	Commission on Innovation and Technology (√)	0.815
Commission	0.222	Workshop	0.078
Innovation and Technology Fund	0.074	CE	0.026
Commission on Innovation & Technology (√)	0.111	Science and Technology	0.026
National Science and Technology Board	0.037	Innovation and Technology Commission (√)	0.026
Innovation and Technology Commission (√)	0.037	Pearl River Delta	0.026
Pearl River Delta	0.037		
Pro-	0.037		
CE	0.037		
Workshop	0.037		
PERSON 陈方安生			
Anson Chan (√)	0.408	Mrs Chan (√)	0.537
Chan (√)	0.343	Anson Chan (√)	0.433
Mrs Chan (√)	0.194	Patrick Lau	0.029
Hon Anson Chan	0.008	Gary Locke	0.029
Hon Mrs Anson Chan	0.006	Progress Report	0.029
Washington	0.003	White House	0.029
Patrick Lau	0.003	……	……
……	……	Administration	0.029
CHAN 1	0.003		
(18 entries)		(12 entries)	

Figure 1: Dictionary sample from the first and last iteration
("(√) " means correct translations)

Baseline annotation

LOCATION{中华人民共和国 香港} 特别 行政区 成立 暨 特区 政府 宣誓就职 仪式 今日 (星期二) 凌晨 在 ORGANIZATION{香港 会议 展览 中心} 新 翼 举行 。

A ceremony to establish the LOCATION{**Hong Kong**} Special ORGANIZATION{**Administrative Region**} (HKSAR) of the LOCATION {**People's Republic of China**} was held early today (Tuesday) at the ORGANIZATION{**Hong Kong Convention and Exhibition Centre (HKCEC**}) Extension .

Annotation after the 5th iteration

LOCATION{中华人民共和国} ORGANIZATION{香港 特别 行政区} 成立 暨 特区 政府 宣誓就职 仪式 今日 (星期二) 凌晨 在 ORGANIZATION{香港 会议 展览 中心} 新 翼 举行 。

A ceremony to establish the LOCATION{**Hong Kong**} Special ORGANIZATION{**Administrative Region**} (ORGANIZATION{**HKSAR**}) of the LOCATION{**People's Republic of China**} was held early today (Tuesday) at the ORGANIZATION{**Hong Kong Convention and Exhibition Centre**} (HKCEC) Extension .

Figure2: Annotation sample from the baseline and 5th iteration

reduced by 51.8% and the annotation quality was improved from 70.03 to 78.15 for Chinese and 73.38 to 81.46 in terms of F-score.

Future work will focus on incorporating the NE detection and translation into the statistical system developed in our group.

7. Acknowledgement

We cordially thank BBN for providing us with their named entity tagging software IdentiFinder™.

8. References

[1] D. Appelt, J. Hobbs, D. Israel, and M. Tyson. Fastus: A finite-state processor for information extraction from real world texts. *In Proceedings of IJCAI-93*, pp.1172-1178, Chambery, France, 1993.

[2] D. Bikel, S. Miller, R. Schwarz and R. Weischedel. Nymble: A high-performance learning name-finder. *In Proceedings of Applied Natural Language Processing,* pp.194-201, Washington DC, 1997.

[3] Y. Al-Onaizan and K. Knight. Named Entity Translation, *in Proceedings of Human Language Technology 2002*, pp.111-115, San Diego, CA, March, 2002.

[4] B. Stalls and K. Knight. Translating Names and Technical Terms in Arabic Text. *In Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, 1998.

[5] H. Meng, W. K. Lo, B. Chen and K. Tang. Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval. *In Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Trento, Italy, December 2001.

[6] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R.L. Mercer. The mathematics of Machine Translation: Parameter Estimation. *In Computational Linguistics,* vol 19, number 2. pp.263-311, June, 1993.

[7] S. Vogel, H. Ney and C. Tillmann. HMM-Based Word Alignment in Statistical Translation. *In Proceedings of the ACL'96*, pp. 836-841, Copenhagen Denmark. August 1996.

[8] I. Dan Melamed. Models of Translational Equivalence among Words, *In Computational Linguistics 26(2),* pp. 221-249, June 2000.

[9] N. Chinchor, P. Robinson and E. Brown. Hub-4 IE-NE Task Definition Version 4.8, http://www.nist.gov/speech/hub4_98/h4_iene_task_def.4.8.ps, August 21, 1998.