Combining Source and Target Language Information for Name Tagging of Machine Translation Output

Shasha Liao
New York University
715 Broadway, 7th floor
New York, NY 10003 USA
liaoss@cs.nyu.edu

Abstract

A Named Entity Recognizer (NER) generally has worse performance on machine translated text, because of the poor syntax of the MT output and other errors in the translation. As some tagging distinctions are clearer in the source, and some in the target, we tried to integrate the tag information from both source and target to improve target language tagging performance, especially recall.

In our experiments with Chinese-to-English MT output, we first used a simple merge of the outputs from an ET (Entity Translation) system and an English NER system, getting an absolute gain of 7.15% in F-measure, from 73.53% to 80.68%. We then trained an MEMM module to integrate them more discriminatively, and got a further average gain of 2.74% in F-measure, from 80.68% to 83.42%.

1 Introduction

Because of the growing multilingual environment for NLP, there is an increasing need to be able to annotate and analyze the output of machine translation (MT) systems. But treating this task as one of processing "ordinary text" can lead to poor results. We examine this problem with respect to the name tagging of English text.

A Named Entity Recognizer (NER) trained on an English corpus does not have the same performance when applied to machine-translated text. From our experiments on NIST 05 Chinese-to-English MT evaluation data, when we used the same English NER to tag the reference translation and the MT output, the F-measure was 81.38% for the reference but only 73.53% for the MT output. There are two primary reasons for this. First, the performance of current translation systems is not very good, and so the output is quite different from Standard English text. The fluency of the translated text will be poor, and the context of a named entity may be weird. Second, the translated text has some foreign names which are hard for the English NER to recognize, even if they are well translated by the MT system, because such names appear very infrequently in the English training corpus.

Training an NER on MT output does not seem to be an attractive solution. It may take a lot of time to manually annotate a large amount of training data, and this labor may have to be repeated for a new MT system or even a new version of an existing MT system. Furthermore, the resulting system may still not work well, in so far as the translation is not good and information is somehow distorted. In fact, sometimes the meanings of the translated sentences are hard to decipher unless we check the source language or get a human translated document as reference. As a result, we need source language information to aid the English NER.

However, it is also not enough to rely entirely on the source language NE results and map them onto the translated English text. First, the word alignment from source language to English generated by the MT system may not be accurate, leading to problems in mapping the Chinese name tags. Second, the translated text is not exactly same as the source language because there may be information missed or added. For example, the Chinese phrase "香港地铁", which is not a name in Chinese, and should be literally translated as

Proceedings of the ACL-08: HLT Student Research Workshop (Companion Volume), pages 19–24,
Columbus, June 2008.

"*the subway in Hong Kong*", may end up being translated to "*mtrc*", the abbreviation of "*The Mass Transit Railway Corporation*", which is an organization in Hong Kong (and so should get a name tag in English).

If we can use the information from both the source language and the translated text, we cannot only find the named entities missed by the English NER, but also modify incorrect boundaries in the English results which are caused by the bad content. However, using word alignment to map the source language information into the English text is problematic, for two reasons: First, the word alignment produced by machine translation is typically not very good, with a Chinese-English AER (alignment error rate) of about 40% (Deng and William 2005). So just using word alignment to map the information would introduce a lot of noise. Second, in the case of function words in English which have no corresponding realization in Chinese, traditional word alignment would align the function word with another Chinese constituent, such as a name, which could lead to boundary errors in tagging English names. We have therefore used an alternative method to fetch the source language information for information extraction, which is called Entity Translation and is described in Section 3.

2 Motivation

When we use the English NER to annotate the translated text, we find that the performance is not as good as English texts. This is due to several types of problems.

2.1 Bad name contexts

Producing correct word order is very hard for a phrase-based MT system, particularly when translating between two such disparate languages, and there are still a lot of Chinese syntax structures left in translated text, which are usually not regular English expressions. As a result, it is hard for the English NER to detect names in these contexts.[1]

> Ex. 1. annan said, "**kumaratunga** president personally against him to areas under guerrilla control field visit because it feared the rebels will use his visit as a political chip"

It is hard to recognize from this example that ***kumaratunga*** is a person name unless we are already familiar with this name or realize this is a normal Chinese expression structure, although not an English one.

> Ex. 2. A reporter from **shantou** <ORG[2]> **university** school of medicine</ORG>, faculty of medicine, **university of** <GPE>**hong kong**</GPE>, <ORG>influenza research center</ORG> was informed that …...

Here source language information can help fix incorrect name boundaries assigned by the English NER, especially from a messy context. In Example 3, the source language tagger can tell us that "shantou university" and "university of hong kong" are two named entities, allowing us to fix the wrong name boundaries of the English NER.

2.2 Bad translations

There are cases where the MT system does not recognize there is a name and translates it as something else, and if we do not refer to the source language, we sometimes cannot understand the sentence, or annotate it.

> Ex. 3. xinhua shanghai , january 1 (<ORG>**feng yizhen su lofty**</ORG>) snow , frozen , and the shanghai airport staff in snow and inalienable .

The translation system does not output the names correctly, and only when we look at the Chinese sentence can we know that there are two person names here, one is "feng yizhen", and the other is "su lofty", where the second one is translated incorrectly. English NER treats the whole as an ORGANIZATION as there is no punctuation to separate the two names.

2.3 Unknown foreign names

There are many Chinese GPE and PERSON names which are missed because they appear rarely in English text, especially city, county or even province names, and so are hard for English NER to detect or classify. However, on the Chinese side, they may be common names and so easily tagged.

[1] The MT system we used generates monocase translations, so we show all the translations in lower case.

[2] We use the entity types of ACE (the Automatic Content Extraction evaluation) for name types. Here ORG = "ORGANIZATION" is the tag for an organization; GPE = "Geo-Political Entity" is the tag for a location with a government; other locations (e.g., "Sahara Desert") are tagged as LOCATION.

Ex. 4. At present, **shishi city** in the province to achieve a village public transportation, village water ; village of cable television .

The city names in examples 4 are famous in Chinese but do not appear much in English text, and so are missed by the English NER; however, a Chinese NER would be able to tag them as named entities.

3 Entity Translation System

The MT pipeline we employ begins with an Entity Translation (ET) system which identifies and translates the names in the text (Heng Ji et al., 2007). This system runs a source-language NER (based on an HMM) and then uses a variety of strategies to translate the names it identifies. One strategy, for example, uses a corpus-trained name transliteration component coupled with a target language model to select the best transliteration. The source text, annotated with name translations, is then passed to a statistical, phrase-based MT system (Zens and Ney, 2004). Depending on its phrase table and language model, this name-aware MT system would decide whether to accept the translation provided by ET. Experiments show that the MT system with ET pre-processing can produce better translations than the MT system alone, with 17% relative error reduction on overall name translation.

The strategy combining multiple transliterations and selection based on a language model is particularly effective for foreign (non-Chinese) person names rendered in Chinese. If these names did not appear in the bilingual training material, they would be mistranslated by an MT system without ET. These names are often also difficult for the English tagger, so ET can benefit both translation and name recognition.

For each name tagged by ET, we see if the translation string proposed by ET appears in the translation produced by the MT system. If so, we use the ET output to assign an ‘ET name type’ to that string in the translation. This approach avoids the problems of using word alignments from the MT system; in particular, the alignment of function words in English with names in Chinese.

4 Integrating source and target information

We first try a very simple merge method to see how much gain can be gotten by simply combining the two sources. After that, we describe a corpus-trained model which addresses some of the tag conflict situations and gets additional gains.

4.1 Results from English NER and ET

First, we analyzed the English NER and ET output to see the named entity distribution of the two sources. We focus on the differences between them because when they agree, we can expect little improvement from using source language information. In the nist05 data, we find 1893 named entities in the English NER output (target language part) and 1968 named entities in the ET output (source language part); 1171 of them are the same. This means that 38.14% of the names tagged in the target language and 40.5% of those in the source language do not have a corresponding tag in the other language, which suggests that the source and target NER may have different strengths on name tagging.

We checked the names which are tagged differently, and there are 347 correct names from ET missed by English NER and 418 from English NER missed by ET.

4.2 Simple Merge

First, in order to see if the ET system can really help the English NER, we do a simple merge experiment, which just adds the named entities extracted from the ET system into the English NER results, so long as there is no conflict between them (i.e., so long as the ET-tagged name does not overlap an English NE-tagged name).

Our experiments show that this simple method can improve the English NER result substantially (Table 5-1), especially for recall, confirming our intuition.

We checked the errors produced by this simple merge method, and divided them into four types.

1. Missed by both sources.
2. Missed by one source and erroneously tagged by the other
3. Erroneously tagged by both sources
4. Conflict situations where the English NE-tagged name is wrong but the ET-tagged name is correct.

Although there is not much we can do for the first three error types, we can address the last error type by some intelligent learning method. In NIST05 data, there are 261 names which have conflicts, and we can get more gains here.

There are two kinds of conflicts: A type conflict which occurs when the ET and English NER tag the same named entity but give it different types; and a boundary conflict which occurs when there is a tag overlap between English NER and ET. We treat these two kinds of conflict differently by using different features to indicate them.

4.3 Maximum Entropy Markov Model

We use a MEMM (Maximum Entropy Markov Model) as our tagging model. An MEMM is a variation on traditional Hidden Markov Models (HMM). Like an HMM, it attempts to characterize a string of tokens as a most likely set of transitions through a Markov model. The MEMM allows observations to be represented as arbitrary overlapping features (such as word, capitalization, formatting, part-of-speech), and defines the conditional probability of state sequences given observation sequences. It does this by using the maximum entropy framework to fit a set of exponential models that represent the probability of a state given an observation and the previous state (McCallum et al. 2000).

In our experiment, we train the maximum entropy framework at the token level, and use the BIO types as the states to be predicted. There are four entity types: PERSON, ORGANIZATION, GPE and LOCATION, and so a total of 9 states.

4.4 Feature Sets for MEMM

In our experiment, we are interested not only in training a module, but also in measuring the different performance for different scales of training corpora. If a small annotated corpus can get reasonable gain, this method for combining taggers will be much more practical.

As a result, we first build a small feature set and enlarge it by adding more features, expecting that the small feature set may get better performance with a small training corpus.

Set 1: Features Focusing on Current Tag and Previous State Information

We first try to use few features to see how much gain we can get if we only consider the tag information from ET and English NER, and the previous state. These features are:

F1: current token's type in ET
F2: current token's type in English NER
F3: Feature1+Feature2
F4: if there is a type conflict + ET type + English NER type
F5: if there is a type conflict +ET type confidence + English NER confidence
F6: if there is a boundary conflict + ET type + English NER type
F7: if there is a boundary conflict + ET token confidence + English NER confidence
F8: state for the previous token

F4 and F5 are used to help resolve the type conflicts, and F6 and F7 to resolve boundary conflicts. When there is a conflict, we need the confidence information from both ET and English NER to indicate which side to choose.

The English NER reports a *margin*, which can be used to gauge tag confidence. The margin is the difference in log probability between the top tagging hypothesis and a hypothesis which assigns the name a different NE tag, or no NE tag. We use this as the confidence of English NER output.

For ET output, the situation is more complicated. We use different confidence methods for type and boundary conflicts. For type conflicts, we use the source of the ET translation as the "type confidence", for example, if the ET result comes from a person name list, the output is probably correct. For boundary conflicts, as the ET system uses some pruning strategy to fix the boundary errors in word alignment, and the translation procedure contains several disparate components which produce different kind of confidence measure, it is not reasonable to use Chinese NER confidence as the confidence estimate. As a result, we check if the token is capitalized in ET translation, and treat it as the "token confidence".

Set 2: Set 1 + Current Token Information

F9: current token + ET type+ English NER type

Token information can be used to predict the result when there is a conflict, as the conflict reason varies and in some cases without knowing the token itself, it is hard to know the right choice. As a result, we add the current token feature but this is the only place we use token information.

Set 3: Set2 + Sequence Information

Our experiments showed some performance gain with only the current token features and the previous state, but we still wanted to see if additional features – such as information on the previous and following tokens – would help. To this end, we added such features, while still retaining our focus on the ET and English NER information:

F10: English NER result of the current token + that of the previous token

F11: ET result of the current token + ET result of the previous token.

F12: English NER result of the current token + that of the next token.

F13: ET result of the current token + that of the next token.

5 Experiment

The experiment was carried out on the Chinese part of the NIST 05 machine translation evaluation (NIST05) and NIST 04 machine translation evaluation (NIST04) data, where NISTT05 contains 100 documents and NIST04 contains 200 documents. We annotated all the data in NIST05 and 120 documents for NIST04 for our experiment.

The ET system used a Chinese HMM-based NER trained on 1,460,648 words; the English name tagger was also HMM-based and trained on 450,000 words.

First, we want to see the result with very small training data, and so divided the NIST05 data into 5 subsets, each containing 20 documents. We ran a cross validation experiment on this small corpus, with 4 subsets as training data and 1 as testing data. We refer to this configuration as Corpus1[3].

Second, to see whether increasing the training data would appreciably influence the result, we added the annotated NIST04 data into the training corpus, and we call this configuration Corpus2.

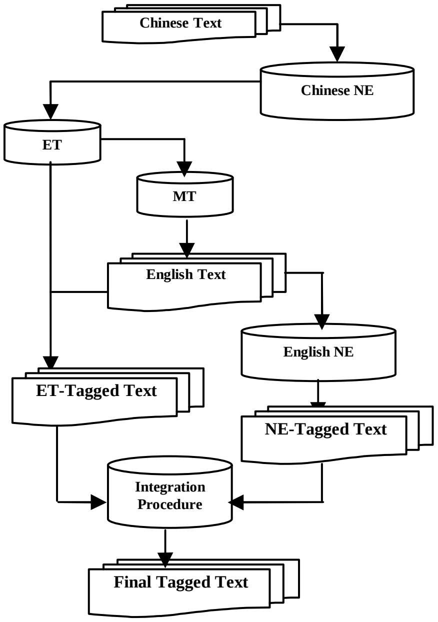

Figure 1. Flow chart of our system

5.1 Simple Merge Result

The simple merge method gets a significant F-measure gain of 7.15% from the English NER baseline, which confirms our intuition that some named entities are easy to tag in source language and others in target language. This represents primarily a significant recall improvement, 14.37%.

	NER baseline	Simple Merge
P	85.68	82.70
R	64.39	78.76
F	73.53	**80.68**

Table 1. Simple merge method on Corpus1 (100 documents)

5.2 Integrating Results on Corpus1

On this small training corpus, we test each subset with other subsets as training data, and calculate the total performance on the whole corpus. The best result comes from Set2 instead of Set3, presumably because the training data is too small to handle the richer model of Set3. Our experiment shows that we can get 1.9% gain over simple merge method with Set 2 using 80 documents as training data.

	Simple Merge	Set1	Set2	Set3
P	82.70	84.73	84.72	84.48
R	78.76	78.01	80.55	80.15
F	80.68	81.23	**82.58**	82.26

[3] We conducted some experiments with a small corpus in which we relied on the alignment information from the MT system, but the results were much worse than using the ET output. Simple merge using alignment yielded a name tagger F score of 73.34% (1.42% worse than the baseline, 75.76%), while ET F score of 81.23%; MEMM with minimal features using alignment yielded an improvement of 1.7% (vs. 7.9% using ET).

Table 2. Results on Corpus1, which contains 100 documents, with 80 documents used for training at each fold.

5.3 Integrating Results on Corpus2

On this corpus, every training data set contains 200 documents, and we can get a gain of 2.74% over the simple merge method. With the larger training set, the richer model (Set 3) now outperforms the others.

	Simple Merge	**Set1**	**Set2**	**Set3**
P	82. 70	85.04	85.15	85.78
R	78. 76	78.09	80.59	81.18
F	80. 68	81.42	82.81	**83.42**

Table 3. Result on Corpus2 (220 documents), with 200 documents used for training at each fold of cross-validation.

On corpus2, Using a Wilcoxon Matched-Pairs test, with a 10-fold division, all the sets perform significantly better (in F-measure) than the simple merge at a 95% confidence level.

6 Prior Work

Huang and Vogel (2002) describe an approach to extract a named entity translation dictionary from a bilingual corpus while concurrently improving the named entity annotation quality. They use a statistical alignment model to align the entities and iteratively extract the name pairs with higher alignment probability and treat them as global information to improve the monolingual named entity annotation quality for both languages. Using this iterative method, they get a smaller but cleaner named entity translation dictionary and improve the annotation F-measure from 70.03 to 78.15 for Chinese and 73.38 to 81.46 in English. This work is similar in using information from the source language (in this case mediated by the word alignment) to improve the target language tagging. However, they used bi-texts (with hand-translated, relatively high-quality English) and so did not encounter the problems, mentioned above, which arise with MT output.

7 Conclusion

We present an integrated approach to extract the named entities from machine translated text, using name entity information from both source and target language. Our experiments show that with a combination of ET and English NER, we can get a considerably better NER result than would be possible with either alone, and in particular, a large improvement in name identification recall.

MT output poses a challenge for any type of language analysis, such as relation or event recognition or predicate-argument analysis. Even though MT is improving, this problem is likely to be with us for some time. The work reported here indicates how source language information can be brought to bear on such tasks.

The best F-measure in our experiments exceeds the score of the English NER on reference text, which reflects the intuition that even for well translated text, we can still benefit from source language information.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency under Contract No. HR0011-06-C-0023, and the National Science Foundation under Grant NO. IIS-0534700. Any opinions, findings and conclusions expressed in this material are those of the author and do not necessarily reflect the views of the U. S. Government.

References

Yonggang Deng, Byrne and William J. 2005. *HMM Word and Phrase Alignment for Statistical Machine Translation.* Proc. Human Language Technology Conference and Empirical Methods in Natural Language Processing**.**

Fei Huang and Vogel, S. 2002. *Improved named entity translation and bilingual named entity extraction.* Proc. Fourth IEEE Int'l. Conf. on Multimodal Interfaces.

A. McCallum, D. Freitag and F. Pereira. 2000. *Maximum entropy Markov models for information extraction and segmentation.* Proc. 17th International Conf. on Machine Learning.

Heng Ji, Matthias Blume, Dayne Freitag,Ralph Grishman, Shahram Khadivi and Richard Zens. 2007. *NYU-Fair Isaac-RWTH Chinese to English Entity Translation 07 System.* Proceedings of ACE ET 2007 PI/Evaluation Workshop. Washington.

Richard Zens and Hermann Ney. 2004. *Improvements in phrase-based statistical Machine Translation.* In Proc. HLT/NAACL, Boston