Name-aware Machine Translation

Haibo Li[†] **Jing Zheng**[‡] **Heng Ji**[†] **Qi Li**[†] **Wen Wang**[‡]

[†] Computer Science Department and Linguistics Department
Queens College and Graduate Center, City University of New York
New York, NY, USA 10016
{lihaibo.c, hengjicuny, liqiearth}@gmail.com

[‡] Speech Technology & Research Laboratory
SRI International
Menlo Park, CA, USA 94025
{zj, wwang}@speech.sri.com

Abstract

We propose a Name-aware Machine Translation (MT) approach which can tightly integrate name processing into MT model, by jointly annotating parallel corpora, extracting name-aware translation grammar and rules, adding name phrase table and name translation driven decoding. Additionally, we also propose a new MT metric to appropriately evaluate the translation quality of informative words, by assigning different weights to different words according to their importance values in a document. Experiments on Chinese-English translation demonstrated the effectiveness of our approach on enhancing the quality of overall translation, name translation and word alignment over a high-quality MT baseline[1].

1 Introduction

A shrinking fraction of the world's Web pages are written in English, therefore the ability to access pages across a range of languages is becoming increasingly important. This need can be addressed in part by cross-lingual information access tasks such as entity linking (McNamee et al., 2011; Cassidy et al., 2012), event extraction (Hakkani-Tur et al., 2007), slot filling (Snover et al., 2011) and question answering (Parton et al., 2009; Parton and McKeown, 2010). A key bottleneck of high-quality cross-lingual information access lies in the performance of Machine Translation (MT). Traditional MT approaches focus on the fluency and accuracy of the overall translation but fall short in their ability to translate certain content words including critical information, especially names. A typical statistical MT system can only translate 60% person names correctly (Ji et al., 2009). Incorrect segmentation and translation of names which often carry central meanings of a sentence can also yield incorrect translation of long contexts. Names have been largely neglected in the prior MT research due to the following reasons:

- The current dominant automatic MT scoring metrics (such as Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002)) treat all words equally, but names have relative low frequency in text (about 6% in newswire and only 3% in web documents) and thus are vastly outnumbered by function words and common nouns, etc..
- Name translations pose a greater complexity because the set of names is open and highly dynamic. It is also important to acknowledge that there are many fundamental differences between the translation of names and other tokens, depending on whether a name is rendered phonetically, semantically, or a mixture of both (Ji et al., 2009).
- The artificial settings of assigning low weights to information translation (compared to overall word translation) in some large-scale government evaluations have discouraged MT developers to spend time and explore resources to tackle this problem.

We propose a novel Name-aware MT (NAMT) approach which can tightly integrate name processing into the training and decoding processes of an end-to-end MT pipeline, and a new name-aware metric to evaluate MT which can assign different weights to different tokens according to their importance values in a document. Compared to previous methods, the novel contributions of our approach are:

1. Tightly integrate joint bilingual name tagging into MT training by coordinating tagged

[1] Some of the resources and open source programs developed in this work are made freely available for research purpose at http://nlp.cs.qc.cuny.edu/NAMT.tgz

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 604–614,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

names in parallel corpora, updating word segmentation, word alignment and grammar extraction (Section 3.1).

2. Tightly integrate name tagging and translation into MT decoding via name-aware grammar (Section 3.2).
3. Optimize name translation and context translation simultaneously and conduct name translation driven decoding with language model (LM) based selection (Section 3.2).
4. Propose a new MT evaluation metric which can discriminate names and non-informative words (Section 4).

2 Baseline MT

As our baseline, we apply a high-performing Chinese-English MT system (Zheng, 2008; Zheng et al., 2009) based on hierarchical phrase-based translation framework (Chiang, 2005). It is based on a weighted synchronous context-free grammar (SCFG). All SCFG rules are associated with a set of features that are used to compute derivation probabilities. The features include:

- Relative frequency in two directions $P(\gamma|\alpha)$ and $P(\alpha|\gamma)$, estimating the likelihoods of one side of the rule r: $X \rightarrow <\gamma, \alpha>$ translating into the other side, where γ and α are strings of terminals and non-terminals in the source side and target side. Non-terminals in γ and α are in one-to-one correspondence.
- Lexical weights in two directions: $P_w(\gamma|\alpha)$ and $P_w(\alpha|\gamma)$, estimating likelihoods of words in one side of the rule r: $X \rightarrow <\gamma, \alpha>$ translating into the other side (Koehn et al., 2003).
- Phrase penalty: a penalty exp(1) for a rule with no non-terminal being used in derivation.
- Rule penalty: a penalty exp(1) for a rule with at least one non-terminal being used in derivation.
- Glue rule penalty: a penalty exp(1) if a glue rule used in derivation.
- Translation length: number of words in translation output.

Our previous work showed that combining multiple LMs trained from different sources can lead to significant improvement. The LM used for decoding is a log-linear combination of four word n-gram LMs which are built on different English corpora (details described in section 5.1), with the LM weights optimized on a development set and determined by minimum error rate training (MERT), to estimate the probability of a word given the preceding words. All four LMs were trained using modified Kneser-Ney smoothing algorithm (Chen and Goodman, 1996) and converted into Bloom filter LMs (Talbot and Brants, 2008) supporting memory map.

The scaling factors for all features are optimized by minimum error rate training algorithm to maximize BLEU score (Och, 2003). Given an input sentence in the source language, translation into the target language is cast as a search problem, where the goal is to find the highest-probability derivation that generates the source-side sentence, using the rules in our SCFG. The source-side derivation corresponds to a synchronous target-side derivation and the terminal yield of this target-side derivation is the output of the system. We employ our CKY-style chart decoder, named SRInterp, to solve the search problem.

3 Name-aware MT

We tightly integrate name processing into the above baseline to construct a NAMT model. Figure 1 depicts the general procedure.

3.1 Training

This basic training process of NAMT requires us to apply a bilingual name tagger to annotate parallel training corpora. Traditional name tagging approaches for single languages cannot address this requirement because they were all built on data and resources which are specific to each language without using any cross-lingual features. In addition, due to separate decoding processes the results on parallel data may not be consistent across languages. We developed a bilingual joint name tagger (Li et al., 2012) based on conditional random fields that incorporates both monolingual and cross-lingual features and conducts joint inference, so that name tagging from two languages can mutually enhance each other and therefore inconsistent results can be corrected simultaneously. This joint name tagger achieved 86.3% bilingual pair F-measure with manual alignment and 84.4% bilingual pair F-measure with automatic alignment as reported in (Li et al., 2012). Given a parallel sentence pair we first apply Giza++ (Och and Ney, 2003) to align words, and apply this join-

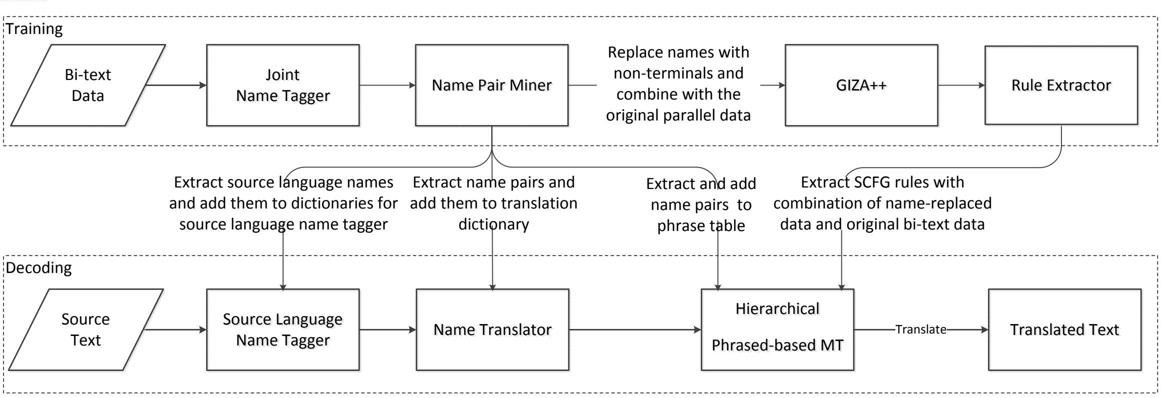
Training
Bi-text Data
Joint Name Tagger
Name Pair Miner
Replace names with non-terminals and combine with the original parallel data
GIZA++
Rule Extractor
Extract source language names and add them to dictionaries for source language name tagger
Extract name pairs and add them to translation dictionary
Extract and add name pairs to phrase table
Extract SCFG rules with combination of name-replaced data and original bi-text data
Decoding
Source Text
Source Language Name Tagger
Name Translator
Hierarchical Phrased-based MT
Translate
Translated Text

Figure 1: Architecture of Name-aware Machine Translation System.

t bilingual name tagger to extract three types of names: (Person (PER), Organization (ORG) and Geo-political entities (GPE)) from both the source side and the target side. We pair two entities from two languages, if they have the same entity type and are mapped together by word alignment. We ignore two kinds of names: multi-word names with conflicting boundaries in two languages and names only identified in one side of a parallel sentence.

We built a NAMT system from such name-tagged parallel corpora. First, we replace tagged name pairs with their entity types, and then use Giza++ and symmetrization heuristics to regenerate word alignment. Since the name tags appear very frequently, the existence of such tags yields improvement in word alignment quality. The re-aligned parallel corpora are used to train our NAMT system based on SCFG. Since the joint name tagger ensures that each tagged source name has a corresponding translation on the target side (and vice versa), we can extract SCFG rules by treating the tagged names as non-terminals.

However, the original parallel corpora contain many high-frequency names, which can already be handled well by the baseline MT. Some of these names carry special meanings that may influence translations of the neighboring words, and thus replacing them with non-terminals can lead to information loss and weaken the translation model. To address this issue, we merged the name-replaced parallel data with the original parallel data and extract grammars from the combined corpus. For example, given the following sentence pair:

- 中国 反对 外来 势力 介入 安哥拉 冲突 .
- China appeals to world for non involvement in Angola conflict .

after name tagging it becomes

- GPE 反对 外来 势力 介入 GPE 冲突 .
- GPE appeals to world for non involvement in GPE conflict .

Both sentence pairs are kept in the combined data to build the translation model.

3.2 Decoding

During decoding phase, we extract names with the baseline monolingual name tagger described in (Li et al., 2012) from a source document. Its performance is comparable to the best reported results on Chinese name tagging on Automatic Content Extraction (ACE) data (Ji and Grishman, 2006; Florian et al., 2006; Zitouni and Florian, 2008; Nguyen et al., 2010). Then we apply a state-of-the-art name translation system (Ji et al., 2009) to translate names into the target language. The name translation system is composed of the following steps: (1) Dictionary matching based on 150,041 name translation pairs; (2) Statistical name transliteration based on a structured perceptron model and a character based MT model (Dayne and Shahram, 2007); (3) Context information extraction based re-ranking.

In our NAMT framework, we add the following extensions to name translation.

We developed a name origin classifier based on Chinese last name list (446 name characters) and name structure parsing features to distinguish Chinese person names and foreign person names (Ji, 2009), so that pinyin conversion is applied for Chinese names while name transliteration is applied only for foreign names. This classifier works reasonably well in most cases (about 92% classification accuracy), except when a common Chinese last name appears as the first character of a foreign name, such as “朱莉” which can be translated either as “*Jolie*” or “*Zhu Li*”.

For those names with fewer than five instances in the training data, we use the name translation system to provide translations; for the rest of the names, we leave them to the baseline MT model to handle. The joint bilingual name tagger was also exploited to mine bilingual name translation pairs from parallel training corpora. The mapping score between a Chinese name and an English name was computed by the number of aligned tokens. A name pair is extracted if the mapping score is the highest among all combinations and the name types on both sides are identical. It is necessary to incorporate word alignment as additional constraints because the order of names is often changed after translation. Finally, the extracted 9,963 unique name translation pairs were also used to create an additional name phrase table for NAMT. Manual evaluation on 2,000 name pairs showed the accuracy is 86%.

The non-terminals in SCFG rules are rewritten to the extracted names during decoding, therefore allow unseen names in the test data to be translated. Finally, based on LMs, our decoder exploits the dynamically created phrase table from name translation, competing with originally extracted rules, to find the best translation for the input sentence.

4 Name-aware MT Evaluation

Traditional MT evaluation metrics such as BLEU (Papineni et al., 2002) and Translation Edit Rate (TER) (Snover et al., 2006) assign the same weights to all tokens equally. For example, incorrect translations of “the” and “Bush” will receive the same penalty. However, for cross-lingual information processing applications, we should acknowledge that certain informationally critical words are more important than other common words. In order to properly evaluate the translation quality of NAMT methods, we propose to modify the BLEU metric so that they can dynamically assign more weights to names during evaluation.

BLEU considers the correspondence between a system translation and a human translation:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

where BP is brevity penalty defined as follows:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ e^{(1-r/c)} & \text{if } c \leq r. \end{cases} \quad (2)$$

where w_n is a set of positive weights summing to one and usually uniformly set as $w_n = 1/N$, c is the length of the system translation and r is the length of reference translation, and p_n is modified n-gram precision defined as:

$$p_n = \frac{\sum\limits_{C \in \text{Candidates}} \sum\limits_{\text{n-gram} \in C} Count_{clip}(\text{n-gram})}{\sum\limits_{C' \in \text{Candidates}} \sum\limits_{\text{n-gram}' \in C'} Count_{clip}(\text{n-gram}')} \quad (3)$$

where C and C' are translation candidates in the candidate sentence set, if a source sentence is translated to many candidate sentences.

As in BLEU metric, we first count the maximum number of times an n-gram occurs in any single reference translation. The total count of each candidate n-gram is clipped at sentence level by its maximum reference count. Then we add up the weights of clipped n-grams and divide them by the total weight of all n-grams.

Based on BLEU score, we design a name-aware BLEU metric as follows. Depending on whether a token t is contained in a name in reference translation, we assign a weight $weight_t$ to t as follows:

$$weight_t = \begin{cases} 1 - e^{-tf(t,d) \cdot idf(t,D)}, & \text{if } t \text{ never appears in names} \\ 1 + \frac{PE}{Z}, & \text{if } t \text{ occurs in name(s)} \end{cases} \quad (4)$$

where PE is the sum of penalties of non-name tokens and Z is the number of tokens within all names:

$$PE = \sum_{t \text{ never appears in names}} e^{-tf(t,d) \cdot idf(t,D)} \quad (5)$$

In this paper, the $tf \cdot idf$ score is computed at sentence level, therefore, D is the sentence set and each $d \in D$ is a sentence.

The weight of an n-gram in reference translation is the sum of weights of all tokens it contains.

$$weight_{ngram} = \sum_{t \in ngram} weight_t \quad (6)$$

Next, we compute the weighted modified n-gram precision $Count_{weight-clip}(\text{n-gram})$ as follows:

$$Count_{weight-clip}(\text{n-gram}) = \sum_{\text{if the } ngram_i \text{ is correctly translated}} weight_{ngram_i} \quad (7)$$

The $Count_{clip}$(n-gram) in the equation 3 is substituted with above $Count_{weight-clip}$(n-gram). When we sum up the total weight of all n-grams of a candidate translation, some n-grams may contain tokens which do not exist in reference translation. We assign the lowest weight of tokens in reference translation to these rare tokens.

We also add an item, name penalty NP, to penalize the output sentences which contain too many or too few names:

$$\text{NP} = e^{-\left(\frac{u}{v}-1\right)^2/2\sigma} \qquad (8)$$

where u is the number of name tokens in system translation and v is the number of name tokens in reference translation.

Finally the name-aware BLEU score is defined as:

$$BLEU_{\text{NA}} = BP \cdot NP \cdot \exp\left(\sum_{n=1}^{N} w_n \log wp_n\right) \qquad (9)$$

This new metric can also be applied to evaluate MT approaches which emphasize other types of facts such as events, by simply replacing name tokens by other fact tokens.

5 Experiments

In this section we present the experimental results of NAMT compared to the baseline MT.

5.1 Data Set

We used a large Chinese-English MT training corpus from various sources and genres (including newswire, web text, broadcast news and broadcast conversations) for our experiments. We also used some translation lexicon data and Wikipedia translations. The majority of the data sets were collected or made available by LDC for U.S. DARPA Translingual Information Detection, Extraction and Summarization (TIDES) program, Global Autonomous Language Exploitation (GALE) program, Broad Operational Language Translation (BOLT) program and National Institute of Standards and Technology (NIST) MT evaluations. The training corpus includes 1,686,458 sentence pairs. The joint name tagger extracted 1,890,335 name pairs (295,087 Persons, 1,269,056 Geopolitical entities and 326,192 Organizations).

Four LMs, denoted LM1, LM2, LM3, and LM4, were trained from different English corpora. LM1 is a 7-gram LM trained on the target side of Chinese-English and Egyptian Arabic-English parallel text, English monolingual discussion forums data R1-R4 released in BOLT Phase 1 (LDC2012E04, LDC2012E16, LDC2012E21, LDC2012E54), and English Gigaword Fifth Edition (LDC2011T07). LM2 is a 7-gram LM trained only on the English monolingual discussion forums data listed above. LM3 is a 4-gram LM trained on the web genre among the target side of all parallel text (i.e., web text from pre-BOLT parallel text and BOLT released discussion forum parallel text). LM4 is a 4-gram LM trained on the English broadcast news and conversation transcripts released under the DARPA GALE program. Note that for LM4 training data, some transcripts were quick transcripts and quick rich transcripts released by LDC, and some were generated by running flexible alignment of closed captions or speech recognition output from LDC on the audio data (Venkataraman et al., 2004).

In order to demonstrate the effectiveness and generality of our approach, we evaluated our approach on seven test sets from multiple genres and domains. We asked four annotators to annotate names in four reference translations of each sentence and an expert annotator to adjudicate results. The detailed statistics and name distribution of each test data set is shown in Table 1. The percentage of names occurred fewer than 5 times in training data are listed in the brackets in the last column of the table.

5.2 Overall Performance

Besides the new name-aware MT metric, we also adopt two traditional metrics, TER to evaluate the overall translation performance and Named Entity Weak Accuracy (NEWA) (Hermjakob et al., 2008) to evaluate the name translation performance.

TER measures the amount of edits required to change a system output into one of the reference translations. Specifically:

$$\text{TER} = \frac{\text{\# of edits}}{\text{average \# of reference words}} \qquad (10)$$

Possible edits include insertion, substitution deletion and shifts of words.

The NEWA metric is defined as follows. Using a manually assembled name variant table, we also support the matching of name variants (e.g., "*World Health Organization*" and "*WHO*").

$$\text{NEWA} = \frac{\text{Count \# of correctly translated names}}{\text{Count \# of names in references}} \qquad (11)$$

Corpus	Genre	Sentence #	Word # in source	Token # in reference	GPE(%)	PER(%)	ORG(%)	All names (% occurred < 5)
BOLT 1	forum	1,200	20,968	24,193	875(82.9)	90(8.5)	91(8.6)	1,056 (51.4)
BOLT 2	forum	1,283	23,707	25,759	815(73.7)	141(12.8)	149(13.5)	1,105 (65.9)
BOLT 3	forum	2,000	38,595	42,519	1,664(80.4)	204(9.8)	204(9.8)	2,072 (47.4)
BOLT 4	forum	1,918	41,759	47,755	1,852(80.0)	348(25.0)	113(5.0)	2,313 (53.3)
BOLT 5	blog	950	23,930	26,875	352(42.5)	235(28.3)	242(29.2)	829 (55.3)
NIST2006	news&blog	1,664	38,442	45,914	1,660(58.2)	568(19.9)	625(21.9)	2,853 (73.1)
NIST2008	news&blog	1,357	32,646	37,315	700(47.9)	367(25.1)	395(27.0)	1,462 (72.0)

Table 1: Statistics and Name Distribution of Test Data Sets.

Metric		System	BOLT 1	BOLT 2	BOLT 3	BOLT 4	BOLT 5	NIST2006	NIST2008
BLEU		Baseline	14.2	14.0	17.3	15.6	15.3	35.5	29.3
		NPhrase	14.1	14.4	17.1	15.4	15.3	35.4	29.3
		NAMT	14.2	**14.6**	16.9	**15.7**	**15.5**	**36.3**	**30.0**
Name-aware BLEU		Baseline	18.2	17.9	18.6	17.6	18.3	36.1	31.7
		NPhrase	18.1	18.8	18.5	18.1	18.0	35.8	31.8
		NAMT	**18.4**	**19.5**	**19.7**	**18.2**	**18.9**	**39.4**	**33.1**
TER		Baseline	70.6	71.0	69.4	70.3	67.1	58.7	61.0
		NPhrase	70.6	70.4	69.4	70.4	67.1	58.7	60.9
		NAMT	**70.3**	**70.2**	**69.2**	**70.1**	**66.6**	**57.7**	**60.5**
NEWA	All	Baseline	69.7	70.1	73.9	72.3	60.6	66.5	60.4
		NPhrase	69.8	71.1	73.8	72.5	60.6	68.3	61.9
		NAMT	**71.4**	**72.0**	**77.7**	**75.1**	**62.7**	**72.9**	**63.2**
	GPE	Baseline	72.8	78.4	80.0	78.7	81.3	79.2	76.0
		NPhrase	73.6	79.3	79.2	78.9	82.3	82.6	79.5
		NAMT	**74.2**	**80.2**	**82.8**	**80.4**	79.3	**85.5**	79.3
	PER	Baseline	53.3	44.7	45.1	49.4	48.9	54.2	51.2
		NPhrase	52.2	45.4	48.9	48.5	47.6	55.1	50.9
		NAMT	**55.6**	**45.4**	**58.8**	**55.2**	**56.2**	**60.0**	**52.3**
	ORG	Baseline	56.0	49.0	52.9	38.1	41.7	44.0	41.3
		NPhrase	50.5	50.3	54.4	40.7	41.3	42.2	40.7
		NAMT	**60.4**	**52.3**	**55.4**	**41.6**	**45.0**	**51.0**	**44.8**

Table 2: Translation Performance (%).

For better comparison with NAMT, besides the original baseline, we develop the other baseline system by adding name translation table into the phrase table (NPhrase).

Table 2 presents the performance of overall translation and name translation. We can see that except for the BOLT3 data set with BLEU metric, our NAMT approach consistently outperformed the baseline system for all data sets with all metrics, and provided up to 23.6% relative error reduction on name translation. According to Wilcoxon Matched-Pairs Signed-Ranks Test, the improvement is not significant with BLEU metric, but is significant at 98% confidence level with all of the other metrics. The gains are more significant for formal genres than informal genres mainly because most of the training data for name tagging and name translation were from newswire. Furthermore, using external name translation table only did not improve translation quality in most test sets except for BOLT2. Therefore, it is important to use name-replaced corpora for rule extraction to fully take advantage of improved word alignment.

Many errors from the baseline MT approach occurred because some parts of out-of-vocabulary names were mistakenly segmented into common words. For example, the baseline MT system mistakenly translated a person name "孙红雷 (*Sun Honglei*)" into "*Sun red thunder*". In informal genres such as discussion forums and web blogs, even common names often appear in rare forms due to misspelling or morphing. For example, "奥霸马 (*Obama*)" was mistakenly translated into "*Ma Olympic*". Such errors can be compounded when word re-ordering was applied. For example, the following sentence: "郭美美的力量还真是强大啊，真是佩服她 (*Guo Meimei's strength really is formidable, I really admire her*)" was mistakenly translated into "*Guo the strength of the America and the America also really strong , ah , really admire her*" by the baseline MT system because the person name "郭美美 (*Guomeimei*)" was mistakenly segmented into three words "郭 (*Guo*)", "美 (*the America*)" and "美 (*the America*)". But our NAMT approach successfully identified and translated this name and also generated better overall translation: "*Guo Meimei 's power is also really strong , ah , really admire her*".

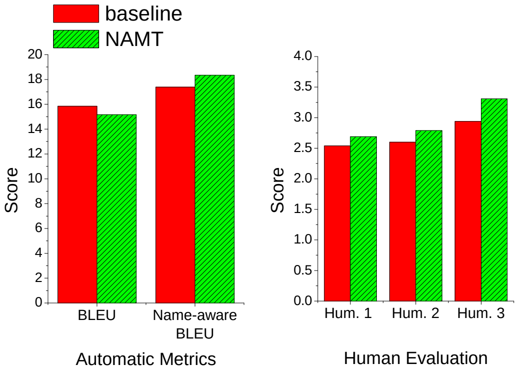

Figure 2: Scores based on Automatic Metrics and Human Evaluation.

5.3 Name-aware BLEU vs The Human Evaluation

In order to investigate the correlation between name-aware BLEU scores and human judgment results, we asked three bi-lingual speakers to judge our translation output from the baseline system and the NAMT system, on a Chinese subset of 250 sentences (each sentence has two corresponding translations from baseline and NAMT) extracted randomly from 7 test corpora. The annotators rated each translation from 1 (very bad) to 5 (very good) and made their judgments based on whether the translation is understandable and conveys the same meaning.

We computed the name-aware BLEU scores on the subset and also the aggregated average scores from human judgments. Figure 2 shows that NAMT consistently achieved higher scores with both name-aware BLEU metric and human judgement. Furthermore, we calculated three Pearson product-moment correlation coefficients between human judgment scores and name-aware BLEU scores of these two MT systems. Give the sample size and the correlation coefficient value, the high significance value of 0.99 indicates that name-aware BLEU tracks human judgment well.

5.4 Word Alignment

It is also important to investigate the impact of our NAMT approach on improving word alignment. We conducted the experiment on the Chinese-English Parallel Treebank (Li et al., 2010) with ground-truth word alignment. The detailed procedure following NAMT framework is as follows: (1) Ran the joint bilingual name tagger; (2) Replaced each name string with its name type (PER, ORG or GPE), and ran Giza++ on the replaced sentences; (3) Ran Giza++ on the words within each name pair. (4) Merged (2) and (3) to produce the final word alignment results. In order to compare with the upper-bound gains, we also measured the performance of applying ground-truth name tagging with the above procedures.

Words	Method	P	R	F
Overall Words	Baseline Giza++	69.8	47.8	56.7
	Joint Name Tagging	**70.4**	**48.1**	**57.1**
	Ground-truth Name Tagging (Upper-bound)	71.3	48.9	58.0
Words Within Names	Baseline Giza++	86.0	31.4	46.0
	Joint Name Tagging	**77.6**	**37.2**	**50.3**

Table 3: Impact of Joint Bilingual Name Tagging on Word Alignment (%).

The experiment results are shown in Table 3. For the words within names, our approach provided significant gains by enhancing F-measure from 46.0% to 50.3%. Only 10.6% words are within names, therefore the upper-bound gains on overall word alignment is only 1.3%. Our joint name tagging approach achieved 0.4% (statistically significant) improvement over the baseline. In Figure 3 we categorized the sentences according to the percentage of name words in each sentence and measured the improvement for each category. We can clearly see that as the sentences include more names, the gains achieved by our approach tend to be greater.

5.5 Remaining Error Analysis

Although the proposed model has significantly enhanced translation quality, some challenges remain. We analyze some major sources of the remaining errors as follows.

1. ***Name Structure Parsing***.

We found that the gains of our NAMT approach were mainly achieved for names with one or two components. When the name structure becomes too complicated to parse, name tagging and name translation are likely to produce errors, especially for long nested organizations. For example, “古田县 检察院 反渎局” (Anti-malfeasance Bureau of Gutian County Procuratorate) consists of a nested organization name with a GPE as modifier: “古田县 检察院” (Gutian County Procuratorate) and an ORG name: “反渎局” (Anti-malfeasance Bureau).

2. ***Name abbreviation tagging and translation***.

Some organization abbreviations are also difficult to extract because our name taggers have

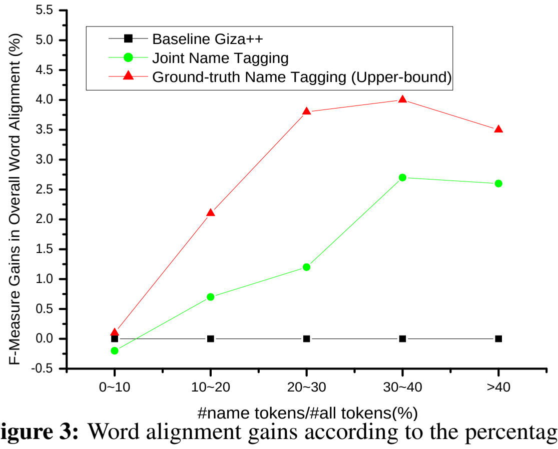

Figure 3: Word alignment gains according to the percentage of name words in each sentence.

not incorporated any coreference resolution techniques. For example, without knowing that "*FAW*" refers to "*First Automotive Works*" in "*FAW has also utilized the capital market to directly finance, and now owns three domestic listed companies*", our system mistakenly labeled it as a GPE. The same challenge exists in name alignment and translation (for example, " 民革 (Min Ge)" refers to " 中国国民党革命委员会" (Revolutionary Committee of the Chinese Kuomintang).

3. ***Cross-lingual information transfer***

English monolingual features normally generate higher confidence than Chinese features for ORG names. On the other hand, some good propagated Chinese features were not able to correct English results. For example, in the following sentence pair: "根据中国，老挝和联合国难民署三方达成的... *(in accordance with the tripartite agreement reached by China, Laos and the UNHCR on)...*", even though the tagger can successfully label "联合国难民署*/UNHCR*" as an organization because it is a common Chinese name, English features based on previous GPE contexts still incorrectly predicted "*UNHCR*" as a GPE name.

6 Related Work

Two types of humble strategies were previously attempted to build name translation components which operate in tandem and loosely integrate into conventional statistical MT systems:

1. *Pre-processing*: identify names in the source texts and propose name translations to the MT system; the name translation results can be simply but aggressively transferred from the source to the target side using word alignment, or added into phrase table in order to enable the LM to decide which translations to choose when encountering the names in the texts (Ji et al., 2009). Heuristic rules or supervised models can be developed to create "do-not-translate" list (Babych and Hartley, 2003) or learn "when-to-transliterate" (Hermjakob et al., 2008).
2. *Post-processing*: in a cross-lingual information retrieval or question answering framework, online query names can be utilized to obtain translation and post-edit MT output (Parton et al., 2009; Ma and McKeown, 2009; Parton and McKeown, 2010; Parton et al., 2012).

It is challenging to decide when to use name translation results. The simple transfer method ensures all name translations appear in the MT output, but it heavily relies on word alignment and does not take into account word re-ordering or the words found in a name's context; therefore it could mistakenly break some context phrase structures due to name translation or alignment errors. The LM selection method often assigns an inappropriate weight to the additional name translation table because it is constructed independently from translation of context words; therefore after weighted voting most correct name translations are not used in the final translation output. Our experimental results 2 confirmed this weakness. More importantly, in these approaches the MT model was still mostly treated as a "black-box" because neither the translation model nor the LM was updated or adapted specifically for names.

Recently the wider idea of incorporating semantics into MT has received increased interests. Most of them designed some certain semantic representations, such as predicate-argument structure or semantic role labeling (Wu and Fung, 2009; Liu and Gildea, 2009; Meyer et al., 2011; Bojar and Wu, 2012), word sense disambiguation (Carpuat and Wu, 2007b; Carpuat and Wu, 2007a) and graph-structured grammar representation (Jones et al., 2012). Lo et al. (2012) proposed a semantic role driven MT metric. However, none of these work declaratively exploited results from information extraction for MT.

Some statistical MT systems (e.g. (Zens et al., 2005), (Aswani and Gaizauskas, 2005)) have attempted to use text normalization to improve word alignment for dates, numbers and job titles. But little reported work has shown the impact of joint name tagging on overall word alignment.

Most of the previous name translation work combined supervised transliteration approaches with LM based re-scoring (Knight and Graehl, 1998; Al-Onaizan and Knight, 2002; Huang et al., 2004). Some recent research used comparable corpora to mine name translation pairs (Feng et al., 2004; Kutsumi et al., 2004; Udupa et al., 2009; Ji, 2009; Fung and Yee, 1998; Rapp, 1999; Shao and Ng, 2004; Lu and Zhao, 2006; Hassan et al., 2007). However, most of these approaches required large amount of seeds, suffered from Information Extraction errors, and relied on phonetic similarity, context co-occurrence and document similarity for re-scoring. In contrast, our name pair mining approach described in this paper does not require any machine translation or transliteration features.

7 Conclusions and Future Work

We developed a name-aware MT framework which tightly integrates name tagging and name translation into training and decoding of MT. Experiments on Chinese-English translation demonstrated the effectiveness of our approach over a high-quality MT baseline in both overall translation and name translation, especially for formal genres. We also proposed a new name-aware evaluation metric. In the future we intend to improve the framework by training a discriminative model to automatically assign weights to combine name translation and baseline translation with additional features including name confidence values, name types and global validation evidence, as well as conducting LM adaptation through bilingual topic modeling and clustering based on name annotations. We also plan to jointly optimize MT and name tagging by propagating multiple word segmentation and name annotation hypotheses in lattice structure to statistical MT and conduct lattice-based decoding (Dyer et al., 2008). Furthermore, we are interested in extending this framework to translate other out-of-vocabulary terms.

Acknowledgement

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF- 09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. NSF EAGER Award under Grant No. IIS-1144111, the U.S. DARPA FA8750-13-2-0041 - Deep Exploration and Filtering of Text (DEFT) Program and CUNY Junior Faculty Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. We express our gratitude to Bing Zhao who provided the test sets and references that were used for Broad Operational Language Translation (BOLT) evaluation and thanks to Taylor Cassidy for constructive comments.

References

Y. Al-Onaizan and K. Knight. 2002. Translating Named Entities Using Monolingual and Bilingual Resources. In *Proceeding ACL'02*, pages 400–408.

N. Aswani and R. Gaizauskas. 2005. A Hybrid Approach to Align Sentences and Words in English-Hindi Parallel Corpora. In *Proceeding ACL'05 Workshop on Building and Using Parallel Texts*, pages 57–64.

Bogdan Babych and Anthony Hartley. 2003. Improving Machine Translation Quality with Automatic Named Entity Recognition. In *Proceeding EAMT '03 workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, pages 1–8.

O. Bojar and D. Wu. 2012. Towards a Predicate-Argument evaluation for MT. In *Proceeding of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 30–38, July.

Marine Carpuat and Dekai Wu. 2007a. How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. In *Proceeding TMI'07*, pages 43–52.

Marine Carpuat and Dekai Wu. 2007b. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceeding EMNLP-CoNLL'07*, pages 61–72.

Taylor Cassidy, Heng Ji, Hongbo Deng, Jing Zheng, and Jiawei Han. 2012. Analysis and Refinement of Cross-lingual Entity Linking. In *Proceeding CLEF'12*, pages 1–12.

Stanley F. Chen and Joshua Goodman. 1996. An Empirical Study of Smoothing Techniques for Language Modeling. *Proceeding of ACL'96*, pages 310–318.

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceeding ACL'05*, pages 263–270.

F. Dayne and K. Shahram. 2007. A Sequence Alignment Model Based on the Averaged Perceptron. In *Proceeding EMNLP-CoNLL'07*, pages 238–247.

C. Dyer, S. Muresan, and P. Resnik. 2008. Generalizing Word Lattice Translation. In *Proceeding ACL-HLT'08*, pages 1012–1020.

D. Feng, Y. Lv, and M. Zhou. 2004. A New Approach for English-Chinese Named Entity Alignment. In *Proceeding PACLIC'04*, pages 372–379.

R. Florian, H. Jing, N. Kambhatla, and I. Zitouni. 2006. Factorizing Complex Models: A Case S-tudy in Mention Detection. In *Proceeding COLING-ACL'06*, pages 473–480.

P. Fung and L. Y. Yee. 1998. An IR Approach for Translating New Words from Nonparallel and Comparable Texts. In *Proceeding COLING-ACL'98*, pages 414–420.

D. Hakkani-Tur, H. Ji, and R. Grishman. 2007. Using Information Extraction to Improve Cross-lingual Document Retrieval. In *Proceeding RANLP Workshop on Multi-source, Multilingual Information Extraction and Summarization*, pages 17–23.

A. Hassan, H. Fahmy, and H. Hassan. 2007. Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora. In *Proceeding RANLP'07*, pages 1–6.

U. Hermjakob, K. Knight, and H. Daume III. 2008. Name Translation in Statistical Machine Translation: Learning When to Transliterate. In *Proceeding ACL'08*, pages 389–397.

F. Huang, S. Vogel, and A. Waibel. 2004. Improving Named Entity Translation Combining Phonetic and Semantic Similarities. In *Proceeding HLT/NAACL'04*, pages 281–288.

H. Ji and R. Grishman. 2006. Analysis and Repair of Name Tagger Errors. In *Proceeding COLING-ACL'06*, pages 420–427.

H. Ji, R. Grishman, D. Freitag, M. Blume, J. Wang, S. Khadivi, R. Zens, and H. Ney. 2009. Name Extraction and Translation for Distillation. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation.*

H. Ji. 2009. Mining Name Translations from Comparable Corpora by Creating Bilingual Information Networks. In *Proceeding ACL-IJCNLP'09 workshop on Building and Using Comparable Corpora*, pages 34–37.

B. Jones, J. Andreas, D. Bauer, K. M. Hermann, and K. Knight. 2012. Semantics-Based Machine Translation with Hyperedge Replacement Grammars. In *Proceeding COLING'12*, pages 1359–1376.

K. Knight and J. Graehl. 1998. Machine Transliteration. In *Computational Linguistics*, volume 24, pages 599–612, Cambridge, MA, USA, December. MIT Press.

P. Koehn, F. Josef Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceeding HLT-NAACL'03*, pages 127–133.

T. Kutsumi, T. Yoshimi, K. Kotani, and I. Sata. 2004. Integrated Use of Internal and External Evidence in The Alignment of Multi-Word Named Entities. In *Proceeding PACLIC'04*, pages 187–196.

X. Li, S. Strassel, S. Grimes, S. Ismael, X. Ma, N. Ge, A. Bies, N. Xue, and M. Maamouri. 2010. Parallel Aligned Treebank Corpora at LDC: Methodology, Annotation and Integration. In *Workshop on Annotation and Exploitation of Parallel Corpora (AEPC).*

Q. Li, H. Li, H. Ji, W. Wang, J. Zheng, and F. Huang. 2012. Joint Bilingual Name Tagging for Parallel Corpora. In *Proceeding CIKM'12*, pages 1727–1731.

D. Liu and D. Gildea. 2009. Semantic Role Features for Machine Translation. In *Proceeding COLING'09*, pages 716–724.

C. Lo, A. K. Tumuluru, and D. Wu. 2012. Fully Automatic Semantic MT Evaluation. In *Proceeding of the Seventh Workshop on Statistical Machine Translation*, pages 243–252.

M. Lu and J. Zhao. 2006. Multi-feature based Chinese-English Named Entity Extraction from Comparable Corpora. In *Proceeding PACLIC'06*, pages 134–141.

W. Ma and K. McKeown. 2009. Where's the Ver-b Correcting Machine Translation During Question Answering. In *Proceeding ACL-IJCNLP'09*, pages 333–336.

P. McNamee, J. Mayfield, D. Lawrie, D. W. Oard, and D. Doermann. 2011. Cross-Language Entity Linking. In *Proceeding IJCNLP'11.*

A. Meyer, M. Kosaka, S. Liao, and N. Xue. 2011. Improving MT Word Alignment Using Aligned Multi-Stage Parses. In *Proceeding ACL-HLT 2011 Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 88–97.

T. T. Nguyen, A. Moschitti, and G. Riccardi. 2010. Kernel-based Reranking for Named-Entity Extraction. In *Proceeding COLING'10*, pages 901–909.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

F. J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceeding A-CL'03*, pages 160–167.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceeding ACL'02*, pages 311–318.

K. Parton and K. McKeown. 2010. MT Error Detection for Cross-Lingual Question Answering. *Proceeding COLING'10*, pages 946–954.

K. Parton, K. R. McKeown, R. Coyne, M. T. Diab, R. Grishman, D. Hakkani-Tur, M. Harper, H. Ji, W. Y. Ma, A. Meyers, S. Stolbach, A. Sun, G. Tur, W. Xu, and S. Yaman. 2009. Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task. In *Proceeding ACL-IJCNLP'09*, pages 423–431.

K. Parton, N. Habash, K. McKeown, G. Iglesias, and A. de Gispert. 2012. Can Automatic Post-Editing Make MT More Meaningful? In *Proceeding EAMT'12*, pages 111–118.

R. Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceeding ACL'99*, pages 519–526.

L. Shao and H. T. Ng. 2004. Mining New Word Translations from Comparable Corpora. In *Proceeding COLING'04*.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceeding of Association for Machine Translation in the Americas*, pages 223–231.

M. Snover, X. Li, W. Lin, Z. Chen, S. Tamang, M. Ge, A. Lee, Q. Li, H. Li, S. Anzaroot, and H. Ji. 2011. Cross-lingual Slot Filling from Comparable Corpora. In *Proceeding ACL'11 Worshop on Building and Using Comparable Corpora*, pages 110–119.

D. Talbot and T. Brants. 2008. Randomized Language Models via Perfect Hash Functions. In *Proceeding of ACL/HLT'08*, pages 505–513.

R. Udupa, K. Saravanan, A. Kumaran, and J. Jagarlamudi. 2009. MINT: A Method for Effective and Scalable Mining of Named Entity Transliterations from Large Comparable Corpora. In *Proceeding EACL'09*, pages 799–807.

A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. R. R. Gadde, and J. Zheng. 2004. An Efficient Repair Procedure For Quick Transcriptions. In *Proceeding INTERSPEECH'04*, pages 1961–1964.

D. Wu and P. Fung. 2009. Semantic Roles for SMT: A Hybrid Two-Pass Model. In *NAACL HLT'09*, pages 13–16.

R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney. 2005. The RWTH Phrase-based Statistical Machine Translation System. In *Proceeding IWSLT'05*, pages 155–162.

J. Zheng, N. F. Ayan, W. Wang, and D. Burkett. 2009. Using Syntax in Large-Scale Audio Document Translation. In *Proceeding Interspeech'09*, pages 440–443.

J. Zheng. 2008. SRInterp: SRI's Scalable Multipurpose SMT Engine. In *Technical Report*.

I. Zitouni and R. Florian. 2008. Mention Detection Crossing the Language Barrier. In *Proceeding EMNLP'08*, pages 600–609.