Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation

Santanu Pal[*], **Sudip Kumar Naskar**[†], **Pavel Pecina**[†],
Sivaji Bandyopadhyay[*] and **Andy Way**[†]

[*]Dept. of Comp. Sc. & Engg.
Jadavpur University
santanupersonal1@gmail.com, sivaji_cse_ju@yahoo.com

[†]CNGL, School of Computing
Dublin City University
{snaskar, ppecina, away}@computing.dcu.ie

Abstract

Data preprocessing plays a crucial role in phrase-based statistical machine translation (PB-SMT). In this paper, we show how single-tokenization of two types of multi-word expressions (MWE), namely named entities (NE) and compound verbs, as well as their prior alignment can boost the performance of PB-SMT. Single-tokenization of compound verbs and named entities (NE) provides significant gains over the baseline PB-SMT system. Automatic alignment of NEs substantially improves the overall MT performance, and thereby the word alignment quality indirectly. For establishing NE alignments, we transliterate source NEs into the target language and then compare them with the target NEs. Target language NEs are first converted into a canonical form before the comparison takes place. Our best system achieves statistically significant improvements (4.59 BLEU points absolute, 52.5% relative improvement) on an English—Bangla translation task.

1 Introduction

Statistical machine translation (SMT) heavily relies on good quality word alignment and phrase alignment tables comprising translation knowledge acquired from a bilingual corpus.

Multi-word expressions (MWE) are defined as “idiosyncratic interpretations that cross word boundaries (or spaces)” (Sag et al., 2002). Traditional approaches to word alignment following IBM Models (Brown et al., 1993) do not work well with multi-word expressions, especially with NEs, due to their inability to handle many-to-many alignments. Firstly, they only carry out alignment between words and do not consider the case of complex expressions, such as multi-word NEs. Secondly, the IBM Models only allow at most one word in the source language to correspond to a word in the target language (Marcu, 2001, Koehn et al., 2003).

In another well-known word alignment approach, Hidden Markov Model (HMM: Vogel et al., 1996), the alignment probabilities depend on the alignment position of the previous word. It does not explicitly consider many-to-many alignment either.

We address this many-to-many alignment problem indirectly. Our objective is to see how to best handle the MWEs in SMT. In this work, two types of MWEs, namely NEs and compound verbs, are automatically identified on both sides of the parallel corpus. Then, source and target language NEs are aligned using a statistical transliteration method. We rely on these automatically aligned NEs and treat them as translation examples. Adding bilingual dictionaries, which in effect are instances of atomic translation pairs, to the parallel corpus is a well-known practice in domain adaptation in SMT (Eck et al., 2004; Wu et al., 2008). We modify the parallel corpus by converting the MWEs into single tokens and adding the aligned NEs in the parallel corpus in a bid to improve the word alignment, and hence the phrase alignment quality. This

Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), pages 45–53,
Beijing, August 2010

preprocessing results in improved MT quality in terms of automatic MT evaluation metrics.

The remainder of the paper is organized as follows. In section 2 we discuss related work. The System is described in Section 3. Section 4 includes the results obtained, together with some analysis. Section 5 concludes, and provides avenues for further work.

2 Related Work

Moore (2003) presented an approach for simultaneous NE identification and translation. He uses capitalization cues for identifying NEs on the English side, and then he applies statistical techniques to decide which portion of the target language corresponds to the specified English NE. Feng et al. (2004) proposed a Maximum Entropy model based approach for English—Chinese NE alignment which significantly outperforms IBM Model4 and HMM. They considered 4 features: translation score, transliteration score, source NE and target NE's co-occurrence score, and the distortion score for distinguishing identical NEs in the same sentence. Huang et al. (2003) proposed a method for automatically extracting NE translingual equivalences between Chinese and English based on multi-feature cost minimization. The costs considered are transliteration cost, word-based translation cost, and NE tagging cost.

Venkatapathy and Joshi (2006) reported a discriminative approach of using the compositionality information about verb-based multi-word expressions to improve word alignment quality. (Ren et al., 2009) presented log likelihood ratio-based hierarchical reducing algorithm to automatically extract bilingual MWEs, and investigated the usefulness of these bilingual MWEs in SMT by integrating bilingual MWEs into Moses (Koehn et al., 2007) in three ways. They observed the highest improvement when they used an additional feature to represent whether or not a bilingual phrase contains bilingual MWEs. This approach was generalized in Carpuat and Diab (2010). In their work, the binary feature was replaced by a count feature representing the number of MWEs in the source language phrase.

Intuitively, MWEs should be both aligned in the parallel corpus and translated as a whole. However, in the state-of-the-art PB-SMT, it could well be the case that constituents of an MWE are marked and aligned as parts of consecutive phrases, since PB-SMT (or any other approaches to SMT) does not generally treat MWEs as special tokens. Another problem SMT suffers from is that verb phrases are often wrongly translated, or even sometimes deleted in the output in order to produce a target sentence considered good by the language model. Moreover, the words inside verb phrases seldom show the tendency of being aligned one-to-one; the alignments of the words inside source and target verb phrases are mostly many-to-many, particularly so for the English—Bangla language pair. These are the motivations behind considering NEs and compound verbs for special treatment in this work.

By converting the MWEs into single tokens, we make sure that PB-SMT also treats them as a whole. The objective of the present work is two-fold; firstly to see how treatment of NEs and compound verbs as a single unit affects the overall MT quality, and secondly whether prior automatic alignment of these single-tokenized MWEs can bring about any further improvement on top of that.

We carried out our experiments on an English—Bangla translation task, a relatively hard task with Bangla being a morphologically richer language.

3 System Description

3.1 PB-SMT

Translation is modeled in SMT as a decision process, in which the translation $e_1^I = e_1 \ldots e_i \ldots e_I$ of a source sentence $f_1^J = f_1 \ldots f_j \ldots f_J$ is chosen to maximize (1):

$$\arg\max_{I,e_1^I} P(e_1^I \mid f_1^J) = \arg\max_{I,e_1^I} P(f_1^J \mid e_1^I).P(e_1^I) \quad (1)$$

where $P(f_1^J \mid e_1^I)$ and $P(e_1^I)$ denote respectively the translation model and the target language model (Brown et al., 1993). In log-linear phrase-based SMT, the posterior probability $P(e_1^I \mid f_1^J)$ is directly modeled as a log-linear combination of features (Och and Ney, 2002), that usually comprise M translational features, and the language model, as in (2):

$$\log P(e_1^I \mid f_1^J) = \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I, s_1^K) + \lambda_{LM} \log P(e_1^I) \qquad (2)$$

where $s_1^k = s_1 ... s_k$ denotes a segmentation of the source and target sentences respectively into the sequences of phrases $(\hat{e}_1,...,\hat{e}_k)$ and $(\hat{f}_1,..., \hat{f}_k)$ such that (we set $i_0 = 0$) (3):

$$\forall 1 \le k \le K, \; s_k = (i_k, b_k, j_k),$$
$$\hat{e}_k = e_{i_{k-1}+1} ... e_{i_k},$$
$$\hat{f}_k = f_{b_k} \cdots f_{j_k}. \qquad (3)$$

and each feature $\hat{h}_m$ in (2) can be rewritten as in (4):

$$h_m(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) \qquad (4)$$

where $\hat{h}_m$ is a feature that applies to a single phrase-pair. It thus follows (5):

$$\sum_{m=1}^{M} \lambda_m \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) = \sum_{k=1}^{K} \hat{h}(\hat{f}_k, \hat{e}_k, s_k) \qquad (5)$$

where $\hat{h} = \sum_{m=1}^{M} \lambda_m \hat{h}_m$.

3.2 Preprocessing of the Parallel Corpus

The initial English—Bangla parallel corpus is cleaned and filtered using a semi-automatic process. We employed two kinds of multi-word information: compound verbs and NEs. Compound verbs are first identified on both sides of the parallel corpus. Chakrabarty et al. (2008) analyzed and identified a category of V+V complex predicates called lexical compound verbs for Hindi. We adapted their strategy for identification of compound verbs in Bangla. In addition to V+V construction, we also consider N+V and ADJ+V structures.

NEs are also identified on both sides of translation pairs. NEs in Bangla are much harder to identify than in English (Ekbal and Bandyopadhyay, 2009). This can be attributed to the fact that (i) there is no concept of capitalization in Bangla; and (ii) Bangla common nouns are often used as proper names. In Bangla, the problem is compounded by the fact that suffixes (case markers, plural markers, emphasizers, specifiers) are also added to proper names, just like to any other common nouns. As a consequence, the accuracy of Bangla NE recognizers (NER) is much poorer compared to that for English. Once the compound verbs and the NEs are identified on both sides of the parallel corpus, they are converted into and replaced by single tokens. When converting these MWEs into single tokens, we replace the spaces with underscores ('_'). Since there are already some hyphenated words in the corpus, we do not use hyphenation for this purpose; besides, the use of a special word separator (underscore in our case) facilitates the job of deciding which single-token (target language) MWEs to detokenize into words comprising them, before evaluation.

3.3 Transliteration Using Modified Joint Source-Channel Model

Li et al. (2004) proposed a generative framework allowing direct orthographical mapping of transliteration units through a joint source-channel model, which is also called n-gram transliteration model. They modeled the segmentation of names into transliteration units (TU) and their alignment preferences using maximum likelihood via EM algorithm (Dempster et al., 1977). Unlike the noisy-channel model, the joint source-channel model tries to capture how source and target names can be generated simultaneously by means of contextual n-grams of the transliteration units. For K aligned TUs, they define the bigram model as in (6):

$$\begin{aligned} P(E,B) &= P(e_1, e_2 ... e_K, b_1, b_2 ... b_K) \\ &= P(<e,b>_1, <e,b>_2 \; ... <e,b>_K) \\ &= \prod_{k=1}^{K} P(<e,b>_k | <e,b>_1^{k-1}) \qquad (6) \end{aligned}$$

where E refers to the English name and B the transliteration in Bengali, while e_i and b_i refer to the i^{th} English and Bangla segment (TU) respectively.

Ekbal et al. (2006) presented a modification to the joint source-channel model to incorporate different contextual information into the model for Indian languages. They used regular expressions and language-specific heuristics based on consonant and vowel patterns to segment names into TUs. Their modified joint source-channel model, for which they obtained improvement over the original joint source-channel model, essentially considers a trigram model for the source language and a bigram model for the target, as in (7).

$$P(E,B)=\prod_{k=1}^{K} P(<e,b>_k|<e,b>_{\text{k-1}},e_{k+1}) \quad (7)$$

Ekbal et al. (2006) reported a word agreement ratio of 67.9% on an English—Bangla transliteration task. In the present work, we use the modified joint source-channel model of (Ekbal et al., 2006) to translate names for establishing NE alignments in the parallel corpus.

3.4 Automatic Alignment of NEs through Transliteration

We first create an NE parallel corpus by extracting the source and target (single token) NEs from the NE-tagged parallel translations in which both sides contain at least one NE. For example, we extract the NE translation pairs given in (9) from the sentence pair shown in (8), where the NEs are shown as italicized.

(8a) *Kirti_Mandir* , where *Mahatma_Gandhi* was born , today houses a photo exhibition on the life and times of the *Mahatma* , a library, a prayer hall and other memorabilia .

(8b) *কির্তী_মন্দির* , যেখানে *মহাত্মা_গান্ধী* জন্মেছিলেন , বর্তমানে সেখানে *মহাত্মার* জীবন ও সেই সময়ের ঘটনাসমূহের একটি চিত্রপ্রদর্শনশালা , একটি লাইব্রেরী ও একটি প্রার্থনা ঘর এবং অন্যান্য স্মৃতিবিজড়িত জিনিসপত্র আছে ।

(9a) Kirti_Mandir Mahatma_Gandhi Mahatma

(9b) কির্তী_মন্দির মহাত্মা_গান্ধী মহাত্মার

Then we try to align the source and target NEs extracted from a parallel sentence, as illustrated in (9). If both sides contain only one NE then the alignment is trivial, and we add such NE pairs to seed another parallel NE corpus that contains examples having only one token in both side. Otherwise, we establish alignments between the source and target NEs using transliteration. We use the joint source-channel model of transliteration (Ekbal et al., 2006) for this purpose.

If both the source and target side contains n number of NEs, and the alignments of n-1 NEs can be established through transliteration or by means of already existing alignments, then the n^{th} alignment is trivial. However, due to the relative performance difference of the NERs for the source and target language, the number of NEs identified on the source and target sides is almost always unequal (see Section 4). Accordingly, we always use transliteration to establish alignments even when it is assumed to be trivial.

Similarly, for multi-word NEs, intra-NE word alignments are established through transliteration or by means of already existing alignments. For a multi-word source NE, if we can align all the words inside the NE with words inside a target NE, then we assume they are translations of each other. Due to the relatively poor performance of the Bangla NER, we also store the immediate left and right neighbouring words for every NE in Bangla, just in case the left or the right word is a valid part of the NE but is not properly tagged by the NER.

As mentioned earlier, since the source side NER is much more reliable than the target side NER, we transliterate the English NEs, and try to align them with the Bangla NEs. For aligning (capitalized) English words to Bangla words, we take the 5 best transliterations produced by the transliteration system for an English word, and compare them against the Bangla words. Bangla NEs often differ in their choice of *matras* (vowel modifiers). Thus we first normalize the Bangla words, both in the target NEs and the transliterated ones, to a canonical form by dropping the matras, and then compare the results. In effect, therefore, we just compare the consonant sequences of every transliteration candidate with that of a target side Bangla word; if they match, then we align the English word with the Bangla word.

নিরজ (ন + ◌ি+ র + জ) -- নীরাজ (ন + ◌ী + র + ◌া + জ)

(10)

The example in (10) illustrates the procedure. Assume, we are trying to align “Niraj” with “নীরাজ”. The transliteration system produces “নিরজ” from the English word “Niraj” and we compare “নিরজ” with “নীরাজ”. Since the consonant sequences match in both words, “নিরজ” is considered a spelling variation of “নীরাজ”, and the English word “Niraj” is aligned to the Bangla word “নীরাজ”.

In this way, we achieve word-level alignments, as well as NE-level alignments. (11) shows the alignments established from (8). The word-level alignments help to establish new word / NE alignments. Word and NE alignments obtained in this way are added to the parallel corpus as additional training data.

(11a) Kirti-Mandir — কির্তী-মন্দির
(11b) Kirti — কির্তী
(11c) Mandir — মন্দির
(11d) Mahatma-Gandhi — মহাত্মা-গান্ধী
(11e) Mahatma — মহাত্মা
(11f) Gandhi — গান্ধী
(11g) Mahatma — মহাত্মার

3.5 Tools and Resources Used

A sentence-aligned English—Bangla parallel corpus containing 14,187 parallel sentences from a travel and tourism domain was used in the present work. The corpus was obtained from the consortium-mode project "Development of English to Indian Languages Machine Translation (EILMT) System"[1].

The Stanford Parser[2] and the CRF chunker[3] were used for identifying compound verbs in the source side of the parallel corpus. The Stanford NER[4] was used to identify NEs on the source side (English) of the parallel corpus.

The sentences on the target side (Bangla) were POS-tagged by using the tools obtained from the consortium mode project "Development of Indian Languages to Indian Languages Machine Translation (ILILMT) System". NEs in Bangla are identified using the NER system of Ekbal and Bandyopadhyay (2008). We use the Stanford Parser, Stanford NER and the NER for Bangla along with the default model files provided, i.e., with no additional training.

The effectiveness of the MWE-aligned parallel corpus developed in the work is demonstrated by using the standard log-linear PB-SMT model as our baseline system: GIZA++ implementation of IBM word alignment model 4, phrase-extraction heuristics described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003) on a held-out development set, target language model with Kneser-Ney smoothing (Kneser and Ney, 1995) trained with SRILM (Stolcke, 2002), and Moses decoder (Koehn et al., 2007).

4 Experiments and Results

We randomly extracted 500 sentences each for the development set and testset from the initial parallel corpus, and treated the rest as the training corpus. After filtering on maximum allowable sentence length of 100 and sentence length ratio of 1:2 (either way), the training corpus contained 13,176 sentences. In addition to the target side of the parallel corpus, a monolingual Bangla corpus containing 293,207 words from the tourism domain was used for the target language model. We experimented with different n-gram settings for the language model and the maximum phrase length, and found that a 4-gram language model and a maximum phrase length of 4 produced the optimum baseline result. We therefore carried out the rest of the experiments using these settings.

In training set	English		Bangla	
	T	U	T	U
Compound verbs	4,874	2,289	14,174	7,154
Single-word NEs	4,720	1,101	5,068	1,175
2-word NEs	4,330	2,961	4,147	3,417
>2 word NEs	1,555	1,271	1,390	1,278
Total NEs	10,605	5,333	10,605	5,870
Total NE words	22,931	8,273	17,107	9,106

Table 1. MWE statistics (T - Total occurrence, U – Unique).

Of the 13,676 sentences in the training and development set, 13,675 sentences had at least one NE on both sides, only 22 sentences had equal number of NEs on both sides, and 13,654 sentences had an unequal number of NEs. Similarly, for the testset, all the sentences had at least one NE on both sides, and none had an equal number of NEs on both sides. It gives an indication of the relative performance differences of the NERs. 6.6% and 6.58% of the source tokens belong to NEs in the training and testset respectively. These statistics reveal the high degree of NEs in the tourism domain data that demands special treatment. Of the 225 unique NEs appearing on the source side of the testset, only 65 NEs are found in the training set.

[1] The EILMT and ILILMT projects are funded by the Department of Information Technology (DIT), Ministry of Communications and Information Technology (MCIT), Government of India.

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] http://crfchunker.sourceforge.net/

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

Experiments		Exp	BLEU	METEOR	NIST	WER	PER	TER
Baseline		1	8.74	20.39	3.98	77.89	62.95	74.60
NEs as Single Tokens (NEaST)	NEs of any length as Single Token (New-MWNEaST)	2	9.15	18.19	3.88	77.81	63.85	74.61
	NEs of length >2 as Single Tokens (MWNE-aST)	3	8.76	18.78	3.86	78.31	63.78	75.15
	2-Word NEs as Single Tokens (2WNEaST)	4	9.13	17.28	3.92	78.12	63.15	74.85
Compound Verbs as Single Tokens (CVaST) †		5	9.56	15.35	3.96	77.60	63.06	74.46
NE Alignment (NEA)	Alignment of NEs of any length (New-MWNEA) †	6	**13.33**	**24.06**	**4.44**	**74.79**	**60.10**	**71.25**
	Alignment of NEs of length upto 2 (New-2WNEA) †	7	10.35	20.93	4.11	76.49	62.20	73.05
	Alignment of NEs of length >2 (MWNEA) †	8	12.39	23.13	4.36	75.51	60.58	72.06
	Alignment of NEs of length 2 (2WNEA) †	9	11.2	23.14	4.26	76.13	60.72	72.57
CVaST +NEaST	New-MWNEaST	10	8.62	16.64	3.73	78.41	65.21	75.47
	MWNEaST	11	8.74	14.68	3.84	78.40	64.05	75.40
	2WNEaST	12	8.85	16.60	3.86	78.17	63.90	75.33
CVaST +NEA	New-MWNEA†	13	11.22	21.02	4.16	75.99	61.96	73.06
	New-2WNEA†	14	10.07	17.67	3.98	77.08	63.35	74.18
	MWNEA†	15	10.34	16.34	4.07	77.12	62.38	73.88
	2WNEA†	16	10.51	18.92	4.08	76.77	62.28	73.56

Table 2. Evaluation results for different experimental setups (The '†' marked systems produce statistically significant improvements on BLEU over the baseline system).

Table 1 shows the MWE statistics of the parallel corpus as identified by the NERs. The average NE length in the training corpus is 2.16 for English and 1.61 for Bangla. As can be seen from Table 1, 44.5% and 47.8% of the NEs are single-word NEs in English and Bangla respectively, which suggests that prior alignment of the single-word NEs, in addition to multi-word NE alignment, should also be beneficial to word and phrase alignment.

Of all the NEs in the training and development sets, the transliteration-based alignment process was able to establish alignments of 4,711 single-word NEs, 4,669 two-word NEs and 1,745 NEs having length more than two. It is to be noted that, some of the single-word NE alignments, as well as two-word NE alignments, result from multi-word NE alignment.

We analyzed the output of the NE alignment module and observed that longer NEs were aligned better than the shorter ones, which is quite intuitive, as longer NEs have more tokens to be considered for intra-NE alignment. Since the NE alignment process is based on transliteration, the alignment method does not work where NEs involve translation or acronyms. We also observed that English multi-word NEs are sometimes fused together into single-word NEs.

We performed three sets of experiments: treating compound verbs as single tokens, treating NEs as single tokens, and the combination thereof. Again for NEs, we carried out three types of preprocessing: single-tokenization of (i) two-word NEs, (ii) more than two-word NEs, and (iii) NEs of any length. We make distinctions among these three to see their relative effects. The development and test sets, as well as the target language monolingual corpus (for language modeling), are also subjected to the same preprocessing of single-tokenizing the MWEs. For NE alignment, we performed experiments using 4 different settings: alignment of (i) NEs of length up to two, (ii) NEs of length two,

(iii) NEs of length greater than two, and (iv) NEs of any length. Before evaluation, the single-token (target language) underscored MWEs are expanded back to words comprising the MWEs.

Since we did not have the gold-standard word alignment, we could not perform intrinsic evaluation of the word alignment. Instead we carry out extrinsic evaluation on the MT quality using the well known automatic MT evaluation metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), NIST (Doddington, 2002), WER, PER and TER (Snover et al., 2006). As can be seen from the evaluation results reported in Table 2, baseline Moses without any preprocessing of the dataset produces a BLEU score of 8.74. The low score can be attributed to the fact that Bangla, a morphologically rich language, is hard to translate into. Moreover, Bangla being a relatively free phrase order language (Ekbal and Bandyopadhyay, 2009) ideally requires multiple set of references for proper evaluation. Hence using a single reference set does not justify evaluating translations in Bangla. Also the training set was not sufficiently large enough for SMT. Treating only longer than 2-word NEs as single tokens does not help improve the overall performance much, while single tokenization of two-word NEs as single tokens produces some improvements (.39 BLEU points absolute, 4.5% relative). Considering compound verbs as single tokens (CVaST) produces a .82 BLEU point improvement (9.4% relative) over the baseline. Strangely, when both compound verbs and NEs together are counted as single tokens, there is hardly any improvement. By contrast, automatic NE alignment (NEA) gives a huge impetus to system performance, the best of them (4.59 BLEU points absolute, 52.5% relative improvement) being the alignment of NEs of any length that produces the best scores across all metrics. When NEA is combined with CVaST, the improvements are substantial, but it can not beat the individual improvement on NEA. The (†) marked systems produce statistically significant improvements as measured by bootstrap resampling method (Koehn, 2004) on BLEU over the baseline system. Metric-wise individual best scores are shown in bold in Table 2.

5 Conclusions and Future Work

In this paper, we have successfully shown how the simple yet effective preprocessing of treating two types of MWEs, namely NEs and compound verbs, as single-tokens, in conjunction with prior NE alignment can boost the performance of PB-SMT system on an English—Bangla translation task. Treating compound verbs as single-tokens provides significant gains over the baseline PB-SMT system. Amongst the MWEs, NEs perhaps play the most important role in MT, as we have clearly demonstrated through experiments that automatic alignment of NEs by means of transliteration improves the overall MT performance substantially across all automatic MT evaluation metrics. Our best system yields 4.59 BLEU points improvement over the baseline, a 52.5% relative increase. We compared a subset of the output of our best system with that of the baseline system, and the output of our best system almost always looks better in terms of either lexical choice or word ordering. The fact that only 28.5% of the testset NEs appear in the training set, yet prior automatic alignment of the NEs brings about so much improvement in terms of MT quality, suggests that it not only improves the NE alignment quality in the phrase table, but word alignment and phrase alignment quality must have also been improved significantly. At the same time, single-tokenization of MWEs makes the dataset sparser, but yet improves the quality of MT output to some extent. Data-driven approaches to MT, specifically for scarce-resource language pairs for which very little parallel texts are available, should benefit from these preprocessing methods. Data sparseness is perhaps the reason why single-tokenization of NEs and compound verbs, both individually and in collaboration, did not add significantly to the scores. However, a significantly large parallel corpus can take care of the data sparseness problem introduced by the single-tokenization of MWEs.

The present work offers several avenues for further work. In future, we will investigate how these automatically aligned NEs can be used as anchor words to directly influence the word alignment process. We will look into whether similar kinds of improvements can be achieved for larger datasets, corpora from different domains and for other language pairs. We will also investigate how NE alignment quality can be improved, especially where NEs involve translation and acronyms. We will also try to perform morphological analysis or stemming on the Bangla side before NE alignment. We will also explore whether discriminative approaches to word alignment can be employed to improve the precision of the NE alignment.

Acknowledgements

This research is partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University, and EU projects PANACEA (Grant 7FP-ITC-248064) and META-NET (Grant FP7-ICT-249119).

References

Banerjee, Satanjeev, and Alon Lavie. 2005. An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pp. 65-72. Ann Arbor, Michigan., pp. 65-72.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19(2):263-311.

Carpuat, Marine, and Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In Proceedings of Human Language Technology conference and the North American Chapter of the Association for Computational Linguistics conference (HLT-NAACL 2010), Los Angeles, CA, pp. 242-245.

Chakrabarti, Debasri, Hemang Mandalia, Ritwik Priya, Vaijayanthi Sarma, and Pushpak Bhattacharyya. 2008. Hindi compound verbs and their automatic extraction. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Posters and demonstrations, Manchester, UK, pp. 27-30.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B (Methodological) 39 (1): 1–38.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In Proceedings of the Second International Conference on Human Language Technology Research (HLT-2002), San Diego, CA, pp. 128-132.

Eck, Matthias, Stephan Vogel, and Alex Waibel. 2004. Improving statistical machine translation in the medical domain using the Unified Medical Language System. In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, pp. 792-798.

Ekbal, Asif, and Sivaji Bandyopadhyay. 2009. Voted NER system using appropriate unlabeled data. In proceedings of the ACL-IJCNLP-2009 Named Entities Workshop (NEWS 2009), Suntec, Singapore, pp. 202-210.

Ekbal, Asif, and Sivaji Bandyopadhyay. 2008. Maximum Entropy Approach for Named Entity Recognition in Indian Languages. International Journal for Computer Processing of Languages (IJCPOL), Vol. 21(3):205-237.

Feng, Donghui, Yajuan Lv, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), Barcelona, Spain, pp. 372-379.

Huang, Fei, Stephan Vogel, and Alex Waibel. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In Proceedings of the ACL-2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, 2003, Sapporo, Japan, pp. 9-16.

Kneser, Reinhard, and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In Proceedings of the IEEE Internation Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 181-184. Detroit, MI.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of HLT-NAACL 2003:

conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series, Edmonton, Canada, pp. 48-54.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL 2007): Proceedings of demo and poster sessions, Prague, Czech Republic, pp. 177-180.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In EMNLP-2004: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 25-26 July 2004, Barcelona, Spain, pp. 388-395.

Marcu, Daniel. 2001. Towards a Unified Approach to Memory- and Statistical-Based Machine Translation. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), Toulouse, France, pp. 386-393.

Moore, Robert C. 2003. Learning translations of named-entity phrases from parallel corpora. In Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), Budapest, Hungary; pp. 259-266.

Och, Franz J. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), Sapporo, Japan, pp. 160-167.

Papineni, Kishore, 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 (ACL-2002), Philadelphia, PA, pp. 311-318.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, Suntec, Singapore, pp. 47-54.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Mexico City, Mexico, pp. 1-15.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006), Cambridge, MA, pp. 223-231.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996), Copenhagen, pp. 836-841.

Venkatapathy, Sriram, and Aravind K. Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In Proceedings of Coling-ACL 2006: Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, pp. 20-27.

Wu, Hua Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), Manchester, UK, pp. 993-1000.