`Manawi`: Using Multi-Word Expressions and Named Entities to Improve Machine Translation

Liling Tan and Santanu Pal

Applied Linguistics, Translation and Interpretation Department
Universität des Saarlandes
`liling.tan@uni-saarland.de`
`santanu.pal@uni-saarland.de`

Abstract

We describe the `Manawi`[1] (मानवि) system submitted to the 2014 WMT translation shared task. We participated in the English-Hindi (EN-HI) and Hindi-English (HI-EN) language pair and achieved 0.792 for the Translation Error Rate (TER) score[2] for EN-HI, the lowest among the competing systems. Our main innovations are (i) the usage of outputs from NLP tools, viz. bilingual multi-word expression extractor and named-entity recognizer to improve SMT quality and (ii) the introduction of a novel filter method based on sentence-alignment features. The `Manawi` system showed the potential of improving translation quality by incorporating multiple NLP tools within the MT pipeline.

1 Introduction

In this paper, we present Saarland University (USAAR) submission to Workshop for Machine Translation 2014 (WMT 2014) using the `Manawi` MT system. We participated in the generic translation shared task for the English-Hindi (EN-HI) and Hindi-English (HI-EN) language pairs.

Our `Manawi` system showcased the incorporation of NLP tools output within the MT pipeline; a bilingual MWE extractor and a bilingual NE recognizer for English and Hindi were implemented. The output from these NLP tools was appended to the training corpus prior to the SMT model training with the MOSES toolkit (Koehn et al., 2007). The resulting system achieves the lowest Translation Error Rate (TER) among competing systems for the English-Hindi language pair.

The rest of the paper is structured as follow: Section 2 describes the implementation of the NLP tools; Section 3 outlines the corpus pre-processing before the MT training process; Section 4 describes the MT system setup; Section 5 describes a simple post-processing component to handle Out-Of-Vocabulary words; Section 6 presents the WMT shared task results for the `Manawi` system and Section 6 concludes the paper.

2 NLP Tools Implementation

2.1 Bilingual MWE in MT

Multi-Word Expressions (MWE) are defined as "idiosyncratic interpretations that cross word boundaries" (Sag et al., 2002). MWE can be made up of collocations (e.g. *seem ridiculous : behuda dikhai*), frozen expressions (e.g. *exception handling : apavada sancalaka*) or name entities (e.g. *Johnny Cash : Johni Kesh*). Jackendoff (1997) claims that the frequency of MWE and the frequency of single words in a speaker's lexicon are almost equivalent.

Bilingual MWE has shown to be useful for a variety of NLP applications such as multilingual information retrieval (Vechtomova, 2005) and Crosslingual/Multilingual Word Sense Disambiguation (Tan and Bond, 2013; Finlayson and Kulkarni, 2011). For machine translation, various studies had introduced bilingual MWE to improve MT system performance. Lambert (2005) introduced bilingual MWE by grouping them as a single token before training alignment models and they showed that it improved alignment and translation quality. Ren et al. (2009) integrated an in-domain bilingual MWE using log likelihood ratio based hierarchical reducing algorithm and gained +0.61 BLEU score. Similarly, Santanu et al. (2010) single tokenized MWE before training a phrase-based SMT model and achieved 50% improvement in BLEU score.

---

[1]**M**ulti-word expression **A**nd **N**amed-entity **A**nd **Wi**kipedia titles (`Manawi`)

[2]Lower TER often results in better translation

In order to improve the word alignment quality, Venkatapathy and Joshi (2006) reported a discriminative approach to use the compositionality information of verb-based multi-word expressions. Pal et al. (2011) discussed the effects of incorporating prior alignment of MWE and NEs directly or indirectly into Phrase-based SMT systems.

2.2 Bilingual MWE Extraction

Monolingual MWE extraction revolves around three approaches (i) rule-based methods relying on morphosyntactic patterns, (ii) statistical methods which use association/frequency measures to determine ngrams as MWE and (iii) hybrid approaches that combine the rule-based and statistical methods.

However, where bilingual MWE extraction techniques are concerned, they operate around two main modus operandi (i) extracting monolingual MWE separately and aligning them at word/phrasal level afterwards or (ii) aligning parallel text at word/phrasal level and then extracting MWE.

We implemented a *language independent bilingual MWE extractor*, (Muwee), that produces a parallel dictionary of MWE *without the need for any word/phrasal-level alignment*. Muwee makes use of the fact that the number of highly collocated MWE should be the same for each sentences pair.

Muwee first extracts MWE separately from the source and target sentences; the MWE are extracted based on bigrams that reports a Pointwise Mutual Information (PMI) score of above 10. Then for each parallel sentence, if the number of MWE are equivalent for the source and target, the bigrams are joint together as a string and contiguous duplicate words are deleted. The removal of contiguous duplicate words is grounded on the fact that linguistically motivated MWE that forms grammatical phrases had shown to improve SMT performances (Pal et al., 2013). Figure 1 presents an example of the MWE extraction process.

MWE with PMI > 10:	['Mahendra Sanskritic', 'Sanskritic University'] ['महेन्द्र संस्कृत', 'संस्कृत बनायागाया']
Concatenated MWE:	'Mahendra Sanskritic Sanskritic University' 'महेन्द्र संस्कृत संस्कृत बनायागाया'
Remove duplicate:	'Mahendra Sanskritic University' 'महेन्द्र संस्कृत बनायागाया'

Figure 1: Muwee Extraction Process

2.3 Named-entity Recognition

Named-Entity (NE) recognition is the task of identifying entities such as names of people, organizations and locations. Given a perfect MWE extraction system, NEs would have been captured by MWE extraction. However, the state-of-art MWE extractors have yet been perfected.

To compliment the MWE extracted by Muwee, we implemented a bilingual NE extractor by combining outputs from the (i) Stanford English NE Recognizer (NER)[3] and (ii) a Do-It-Yourself (DIY) Hindi NER using CRF++ toolkit[4] with annotated data from NER-SSEA 2008 shared task (Rajeev Sangal and Singh, 2008). We trained a Conditional Random Field classifier for the Hindi NER using unigram features, bigram features and a context window of two words to the left and to the right. And we used the DIY Hindi NER and Stanford NER tool to monolingually annotate the NEs from training corpus for the EN-HI / HI-EN language pair.

Similar to the Muwee bilingual extraction criteria, if the number of NEs are the same on the source and target language, the NEs were joint together as a string. We note that sometimes the bilingual NER output contains more than one NE per sentence. For example, our bilingual NER extractor outputs "*Kalpna Chawla Gurdeep Pandher*", which contains two NEs '*Kalpna Chawla*' and '*Gurdeep Pandher*'. Although the resulting bilingual NE does not provide a perfect NE dictionary, it filters out NEs from the sentence and improves word alignments at the start of the MT pipeline.

3 Corpus Preprocessing

The performance of any data driven SMT depends on the quality of training data. Previous studies had shown that filtering out low quality sentence pairs improves the quality of machine translation. For instance, the Moore-Lewis filter removes sentence pairs based on source-side cross-entropy differences (Moore and Lewis, 2010) and the Edinburgh's MT system used the Modified Moore-Lewis filtering (Axelrod et al., 2011) in WMT 2013 shared task (Durrani et al., 2013). CNGL-DCU system extended the Moore-Lewis filter by incorporating lemmas and named enti-

---

[3]http://nlp.stanford.edu/software/CRF-NER.shtml
[4]http://crfpp.googlecode.com

ties in their definition of perplexity[5] (Rubino et al., 2013; Toral, 2013).

The RWTH Aachen system filtered the Common Crawl Corpus by keeping only sentence pairs that contains at least 70% of the word from a known vocabulary dataset extracted from the other corpora in the WMT 2013 shared task (Peitz et al., 2013). The Docent system from Uppsala University also performed data cleaning on the Common Crawl dataset prior to SMT but they were using more aggressive conditions by (i) removing documents that were identified correctly using a language identification module and (ii) removing documents that falls below a threshold value of alignment points and sentence length ratio (Stymne et al., 2013). Our approach to data cleaning is similar to the Uppsala's system but instead of capitalizing on word-alignments features, we were cleaning the data based on sentence alignment features.

3.1 `GaCha` Filtering: Filter by Character Mean Ratio

Stymne et al. (2013) improved translation quality by cleaning the Common Crawl corpus during the WMT 2013 shared task. They filtered out documents exceeding 60 words and cleaned the remainder of the corpus by exploiting the number of alignment points in word alignments between sentence pairs. Their hypothesis was that sentence pairs with very few alignment points in the intersection would mostly likely not be parallel. This is based on the fact that when using GIZA++ (Och and Ney, 2003), the intersection of alignments is more sparse than the standard SMT symmetrization heuristics like grow-diag-final-and (Koehn, 2005).

Different from Stymne et al., our hypothesis for non-parallelness adheres to sentence level alignment criteria as defined in the Gale-Church algorithm (Gale and Church, 1993). If a sentence pair is parallel, the ratio of the number of characters in the source and target sentence should be coherent to the global ratio of the number of source-target characters in a fully parallel corpus. The Gale-Church algorithm had its parameters tuned to suit European languages and Tan (2013) had demonstrated that sentence-level alignments can be improved by using corpus specific parameters. When

using variable parameters to the Gale-Church algorithm, Tan showed that instead of the default parameters set in the original Gale-Church algorithm, using mean ratio of the noisy corpus can also improve sentence level alignments although the ratio from a clean corpus would achieve even better alignments.

Given the premises of the sentence level alignment hypothesis, we clean the training corpus by first calculating the global mean ratio of the number of characters of source sentence to target sentence and then filter out sentence pairs that exceeds or fall below 20% of the global ratio. We call this method, `GaCha` filtering; this cleaning method is more aggressive than cleaning methods described by Stymne et al. but it filters out noisy sentence level alignments created by non-language specific parameters used by sentence aligners such as Gale-Church algorithm.

3.2 Filtering Noise in HindEnCorp

After manual inspection 100 random sentence pairs from the HindEnCorp (Bojar et al., 2014), we found that documents were often misaligned at sentence level or contains HTML special characters. To further reduce the noise in the HindEnCorp, the `Manawi` system was only trained a subset of the HindEnCorp from the following sources (i) DanielPipes, (ii) TIDES and (iii) EILMT. Lastly, we filtered the training data on allowing a maximum of 100 tokens per language per sentence.

Finally, the cleaned data contained 87,692 sentences, only ∼36% of the original HindEnCorp training data.

4 System Setup

Data: To train the baseline translation model, we have used the cleaned subset of the data as described in Section 3. For the `Manawi` model, we added the NLP outputs from the MWE and NE extractors presented in Section 2. To train the monolingual language model, we used the Hindi sentences from the HindEnCorp.

System: We used the standard log-linear Phrase based SMT model provided from the MOSES toolkit.

Configuration: We experimented with various maximum phrase length for the translation and n-

---

[5]The exponent of cross-entropy may be regarded as perplexity

Manawi Submissions (EN-HI)	BLEU	BLEU (cased)	TER
PB-SMT + MWE + NE	**9.9**	7.1	0.869
PB-SMT + MWE + NE + Wiki (`Manawi`)	7.7	7.6	0.864
`Manawi` + `GaCha` Filter	8.9	**8.9**	0.818
`Manawi` + `GaCha` Filter + Handle OOV	8.8	8.8	0.800
`Manawi` + `GaCha` Filter + Remove OOV	8,9	8.8	**0.792**

Table 1: `Manawi` System Submissions @ WMT 2014 Translation Shared Task for English-Hindi

Manawi Submissions (HI-EN)	BLEU	BLEU (cased)	TER
PB-SMT + MWE + NE + Wiki (`Manawi`)	7.7	7.6	0.864
`Manawi` + `GaCha` Filter	**8.9**	**8.9**	**0.818**

Table 2: `Manawi` System Submissions @ WMT 2014 Translation Shared Task for Hindi-English

gram settings for the language model. And we found that using a *maximum phrase length of 5* and *4-gram language model* produced best result in terms of BLEU and TER for our baseline model (i.e. without the incorporation of outputs from the NLP tools). The other experimental settings were:

- *GIZA++* implementation of IBM word alignment model 4 with grow-diagonal-final-and heuristics for performing word alignment and phrase-extraction (Koehn et al., 2003)

- *Minimum Error Rate Training (MERT)* (Och, 2003) on a held-out development set, target language model with Kneser-Ney smoothing (Kneser and Ney, 1995) using language models trained with SRILM (Stolcke, 2002)

- Reordering model[6] was trained on bidirectional (i.e. using both forward and backward models) and conditioned on both source and target language. The reordering model is built by calculating the probabilities of the phrase pair being associated with the given orientation.

Innovation: We demonstrated the incorporation of multiple NLP tools outputs in the SMT pipline by simply using automatically extracted bilingual MWE and NEs as additional parallel data to the cleaned data and ran the translation and statistical model as per the baseline configurations.

---

[6]For reordering we used lexicalized reordering model, which consists of three different types of reordering by conditioning the orientation of previous and next phrases-monotone (m), swap (s) and discontinuous (d).

5 Post-processing

The MOSES decoder produces translations with Out-Of-Vocabulary (OOV) words that were not translated from the source language. The `Manawi` system post-processed the decoder output by (i) *handling OOV words* by replacing each OOV word with the most probable translation using the lexical files generated by GIZA++ and (ii) *removing OOV words* from the decoded outputs.

6 Results

Table 1 summarizes the `Manawi` system submissions for the English-Hindi language pair for WMT 2014 generic translation shared task. The basic `Manawi` system is a Phrase-based SMT (PB-SMT) setup using extracted MWE and NEs and Wikipedia titles as additional parallel data (i.e. PB-SMT+MWE+NE+Wiki in Table 1). The basic `Manawi` system achieved 7.7 BLEU score and 0.864 TER.

After filtering the data before training the translation model, the `Manawi` system performed better at 8.9 BLEU and 0.818 TER. By adding the post-processing component, we achieved the lowest TER score among competing team at 0.792.

7 Conclusion

The `Manawi` system showed how simple yet effective pre-processing and integration of output from NLP tools improves the performance of MT systems. Using `GaCha` filtering to remove noisy data and using automatically extracted MWE and NEs as additional parallel data improve word and phrasal alignments at the start of the MT pipeline

which eventually improves the quality of machine translation. The best setup for the `Manawi` system achieved the best TER score among the competing system.

Also, the incremental improvements made by step-wise implementation of (i) filtering, (ii) incorporating outputs from NLP tools and (iii) post-processing showed that individual components of the `Manawi` can be integrated into other MT systems without detrimental effects.

Acknowledgments

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Aleš Tamchyna, and Dan Zeman. 2014. Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC'14)*, Reykjavik, Iceland, may. ELRA, European Language Resources Association. in prep.

Nadir Durrani, Barry Haddow, Kenneth Heafield, and Philipp Koehn. 2013. Edinburghs machine translation systems for european language pairs. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 112–119.

Mark Alan Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, MWE '11, pages 20–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

William A Gale and Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5:79–86.

Patrik Lambert. 2005. Data inferred multi-word expressions for statistical machine translation. In *In MT Summit X*.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Santanu Pal, Tanmoy Chakraborty, and Sivaji Bandyopadhyay. 2011. Handling multiword expressions in phrase-based statistical machine translation. In *In Proceedings of the 13th Machine Translation Summit*, pages 215–224. MT Summit 2011.

Santanu Pal, Mahammed Hasanuzzaman, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013. Impact of linguistically motivated shallow phrases in pb-smt. In *ICON 2013 http://sivajibandyopadhyay.com/publications/Icon-v1.3-camera.pdf*. ICON 2013.

Stephan Peitz, Jan-Thorsten Peter Saab Mansour, Christoph Schmidt, Joern Wuebker, Matthias Huck, Markus Freitag, and Hermann Ney. 2013. The rwth aachen machine translation system for wmt 2013. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 191–197.

Dipti Misra Sharma Rajeev Sangal and Anil Kumar Singh, editors. 2008. *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*. Asian Federation of Natural Language Processing, Hyderabad, India, January.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raphael Rubino, Antonio Toral, S Cortés Vaıllo, Jun Xie, Xiaofeng Wu, Stephen Doherty, and Qun Liu. 2013. The cngl-dcu-prompsit translation systems for wmt13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 211–216.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer Berlin Heidelberg.

Pal Santanu, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay, and Andy Way. 2010. Handling named entities and compound verbs in phrase-based statistical machine translation. In *23rd International Conference of Computational Linguistics (Coling 2010), Beijing, Chaina*, pages 46–54.

Sara Stymne, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. Tunable distortion limits and corpus cleaning for smt. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 225–231.

Liling Tan and Francis Bond. 2013. Xling: Matching query sentences to a parallel corpus using topic models for word sense disambiguation.

Liling Tan. 2013. Gachalign: Gale-church sentence-level alignments with variable parameters [software]. Retrieved from https://db.tt/LLrul4zP and https://code.google.com/p/gachalign/.

Antonio Toral. 2013. Hybrid selection of language model training data using linguistic information and perplexity. *ACL 2013*, page 8.

Olga Vechtomova. 2005. The role of multi-word units in interactive information retrieval. In *ECIR*, pages 403–420.

Sriram Venkatapathy and Aravind K Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20–27. Association for Computational Linguistics.