Reducing Asymmetry between language-pairs to Improve Alignment and Translation Quality

Rashmi Gangadharaiah
IBM Research, India
rashgang@in.ibm.com

Abstract

This paper presents a novel method to remove asymmetry between the source and the target languages thereby improving alignment and machine translation (MT) quality. Some words in the source language are redundant for MT tasks but necessary for the source sentence to be grammatical. This paper proposes a method to automatically detect such words. In addition, constraints under which these words should or should not be removed are extracted automatically from the target language. A lattice scheme is used for test sentences to provide alternate paths with and without removal of these words. Such a constraint-based removal technique gives a significant improvement ($p < 0.001$) of 5.29 BLEU points over the baseline Phrase-based MT system for the English-Hindi language-pair.

1 Introduction

Different languages express the same piece of information with different number of words. As a result, in many language-pairs, not all source words have correspondences in the target half of the sentence-pair. The aligner is expected to align these redundant source words to an empty word ("*NULL*") in the target sentence. However, in data-sparse conditions, this is not perfectly learnt.

The IBM models in *GIZA++* (Och and Ney, 2003) align each source word to exactly one target word and a target word can be aligned to multiple source words. Currently in most MT systems, this limitation in alignment is overcome by aligning the data bi-directionally (Koehn et al., 2003) and later combining the resulting alignments. Inspite of using information from bi-directional alignments, due to the noisy nature and limitations in the amount of parallel data available, low quality word-alignments are obtained. To illustrate this, consider the top ten maximum likelihood lexical translation-table entries for the English source words, "*the*" and "*india*" obtained after combining the bi-directional alignments with the English-Hindi language-pair in Figures 3A and 3B. For the sake of clarity, the actual English translations of the Hindi words (with case-markers transliterated) are displayed instead of the Hindi words themselves. As seen, the probability mass is used up by many other Hindi words that are not the right translations of the source words. Also, in Figure 3B, "*india*" has a higher probability of being aligned to "*NULL*" than to its actual translation. The most common errors in data that result in low quality alignments and translations are:

1) Property of a language-pair: Certain words are necessary to be present in a sentence for the sentence to be grammatical but unnecessary for MT. For example, *the* in *"the government was highly perturbed by his activities"* has no correspondence in Hindi. For MT, these words can hurt the performance of the aligner (illustrated in Figure 1) and translation quality (Section 4). This is true for many language-pairs including English to all Indian languages.

2) Noisy and Imperfect nature of the data: (a) Human translation errors: Humans cause typograpical errors (many times the same error is created consistently) during translation. Usually monolingual data is first collected from a source such as the world-wide-web and then given to human translators who are usually non-experts in MT and add redundant words or leave certain source words untranslated.

(b) Errors in automatic extraction: the world-wide-web is often used to obtain parallel data. Articles describing the same event in multiple languages are aligned at the sentence-level to create parallel corpora. Sentence-alignment is not

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1346–1350,
Chiang Mai, Thailand, November 8 – 13, 2011.

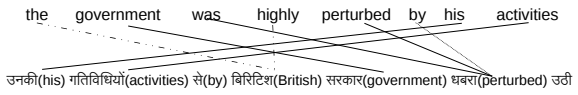

Figure 1: Alignment links provided by the aligner. Dotted lines indicate alignment errors. Correct translations of the Hindi words given in brackets.

perfect when the data used to train the sentence-aligner is limited. Also, not all words in a source sentence have actual correspondences in the target (and vice versa). Errors in word-alignments are seen when a word-aligner is forced to align words in this noisy corpus. For example, the Hindi sentence in Figure 1 has a word that corresponds to "*British*" which is absent in the English sentence.

2 Related Work

Lee (2004) induced only morphological symmetry by identifying morphemes to be merged or deleted from their stems in the morphologically rich source language. They implement the idea that if a morpheme in the source language is robustly translated into a distinct POS in the other language, the morpheme is likely to have an independent counterpart in the other language.

Since unaligned words cause ambiguity in phrase-pairs, Zhang et al. (2009) remove unaligned words before extracting phrase-pairs. Candidates for deletion are collected based on their POS tags and a simple threshold scheme on the probability of a word being aligned.

Lee et al. (2006) found many unaligned function words while translating from Korean to English and removed them using their POS information. In Li et al. (2008), source words that do not have an alignment are added to the phrase-table with ϵ as their translation. Only source context and POS features are used to determine if a word is spurious. For many language-pairs including languages considered in this paper, source context and POS tags of words do not always indicate their 'spuriousness'. Hence, we need other features.

All the above methods depend on a POS tagger to collect redundant source words which may not be available for rare and low density languages. Li et al. (2008) and Zhang et al. (2009) only modify the phrase-extractor or the phrase-pairs and do not concentrate on improving the word-alignments between the language-pair. As will be shown in this paper, the information that a source word is spurious can be valuable even for word-alignment which in turn improves translation quality.

Hong et al. (2008) insert pseudo words and reorder the source and the target sentences to have similar lengths. Pseudo words are generated with the help of dependency parsers. Chung and Gildea (2010) look at recovering dropped pronouns while translating from pro-drop languages like, Chinese and Korean to English.

In this paper, frequently occuring redundant source words caused due to the inherent nature of the language-pair or noise are automatically detected and removed to improve alignments of other words in the corpus and ultimately translation scores when data available is limited. The method does not make use of any rich knowledge sources. Features extracted from word-alignment models are used to score and rank words that form candidates for removal. Top ranking N Redundant source words satisfying target constraints are removed from the training corpus and the corpus is re-aligned. Redundancy is not removed in the target as it hinders the translation generation process and would require post-processing of the translations. During testing, a lattice scheme is used to provide alternate paths, both with and without removal of redundant words.

3 Method for detecting redundant words

Bi-directional word-alignments of the training data are first obtained using GIZA++ and later combined using *grow-diag-final-and* in Moses (Koehn et al., 2007). A maximum likelihood lexical translation table $p(w_i|w_j)$ is estimated between the source words (w_j) and target words (w_i) from the alignments. Features are extracted from the resultant lexical probabilities and linearly combined to obtain a score (eqn. 1). The weights can be tuned with the objective of improving the alignment (with hand alignments) or the translation quality on a tuning set. This can be done using a simple hill climbing procedure or any unconstrained optimization technique that does not require derivatives (Powell, 1964).

$$score = \sum_i \lambda_i f_i \quad where, \sum_i \lambda_i = 1 \qquad (1)$$

3.1 Features

Entropy: If the distribution of the lexical translation probabilities ($P(w_i|w_j)$) for a source word, w_j, (excluding its "*NULL*" probability) is close to

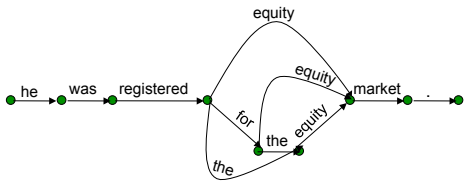

Figure 2: Test/input sentences as a lattice.

a uniform distribution, the entropies seen will be higher than seen with a non-uniform distribution. For example, English words such as "*the*" do not have translations in Indian languages. A plot of the top ten lexical translation probabilities and corresponding lexical translations for "*the*" is given in Figure 3A. Since the "*NULL*" alignment probability is not considered in eqn. 2, the remaining lexical translation probabilities for a source word are re-normalized to sum to 1.

$$f_1(w_j) = \sum_i -p(w_i|w_j)logp(w_i|w_j) \quad (2)$$

Probability of corresponding to "*NULL*": A very high probability of being aligned to "*NULL*" ($f_2(w_j) = p(NULL|w_j)$) is a good indicator that w_j has no target correspondences.

Number of unique target words: The number of lexical translation probability entries for w_j,

$$f_3(w_j) = \frac{\#unique\ translations\ for\ w_j}{\#target\ words\ in\ the\ corpus} \quad (3)$$

Constraints for removal of source words: certain words like "*a*" or "*an*" in English do not always have translations in Hindi. When these words do have a translation, we do not want to remove them from the source sentence. Hence, constraints are extracted from the lexical translation table to determine when a word from a source sentence can be removed. Figure 3 shows the top ranking lexical translations for the source words, "*the*" and "*india*". Target words (w_i) for a given source word (w_j) with translation probability ($p(w_i|w_j)$) less than the thresholds in eqn. 4 are removed from the constraint list of w_j. The thresholds are determined from all the translation probabilities for a given source word (w_j). The ratio between the number of target words removed to the number of words the source word was originally aligned to is also considered as a feature ($f_4(w_j)$).

$$th_1(w_j) = median[p(w_1|w_j), p(w_2|w_j)...] \&\&$$
$$th_2(w_j) = 0.7 * max[p(w_1|w_j), p(w_2|w_j)...] \quad (4)$$

a		the		india	
एक	(one/a)	के	(of/CM)	भारत	(india)
कौ	(of/CM)	की	(of/CM)	इंडिया	(india)
में	(in/CM)	को	(to/CM)	भारतीय	(indian)
का	(of/CM)	का	(of/CM)	हिंदुस्तान	(india)
कुछ	(some/anything)	में	(in/CM)	देश	(country)
कोई	(someone)	ने	(CM)	निगम	(corporation)
किसी	(some)	यह	(this)	हिन्दूस्तानी	(indian)
यह	(this/it)	इस	(this)	अखिल	(all)
(		पर	(but)	आकाशवाणी	(air)
इस	(this)	इन	(these)		
		एक	(one/a)		

Table 1: Constraints for English source words, "*a*", "*the*" and "*india*".

Top ranking N redundant candidates are removed from the source half of the training corpus based on their target constraints. However, during testing, as the reference translations cannot be used, test sentences are converted into a lattice (Dyer et al., 2008) where two alternate paths are included, one with the redundant source word removed and another with the redundant source word as is. An example input lattice for the test sentence "*he was registered for the equity market .*" is given in Figure 2. The scores for taking each of the paths are computed as follows. For any redundant source word, the ratio between the number of times the source word was removed from the training corpus to the total number of times it appeared in the training corpus ($score_1$) is used to score the path that does not include the source word. ($score_2 = 1 - score_1$) is added to the alternate path. If a path contains multiple source words removed, the products of the probabilities ($score_1$) of the redundant source words is taken.

4 Preliminary Results and Analysis

The experiments in this paper are performed with the English-Hindi language-pair. We chose Hindi as it is one of the few Indian languages that have moderate amounts of data to perform translation tasks. However, the method adopted and the analysis done in this paper can be applied to all Indian languages. Indian languages not only suffer from data sparsity, many of these languages also do not have rich knowledge sources (like, POS taggers, parsers, etc.). The parallel corpus included the Darpa TIDES surprise language dataset in 2002 and internally collected parallel corpus from news articles. 200k sentence-pairs were used for training both the Baseline (no removal of redundant source words) and the system with removal of redundant source words using Moses. The basic parameters of Moses were tuned using MERT (Och,

f_1	f_2	f_3	f_4	$comb$
the	-	the	it	the
there	‘s	in	when	in
a	,	a	i	a
says	has	it	but	“
“	will	there	this	as
in	.	“	one	for
what	country	as	some	with
as	have	for	the	to
after	is	this	he	an
for	of	to	if	there

Table 2: Top ranking candidates for removal sorted with respect to each of the feature scores.

2003) on a development set of 500 sentences. The test set contained 4000 sentences. The test set was divided into 20 sub files to determine the statistical significance with the Wilcoxon signed-rank test (1945).

Features from Section 3.1 are normalized to fall within [0,1]. The weights are tuned using a simple grid-search that tried multiple combinations of weights, starting from 0 and keeping the sum of the weights equal to 1. The weights were incremented in steps of 0.2. The objective chosen was to improve the translation quality score (BLEU (Papineni et al., 2002)) on a small tune set of 200 sentence-pairs. For each set of parameters, the training corpus was modified by removing top ranking N redundant candidates based on target constraints and aligned, the tune set was also modified each time into a lattice and translated to compute the translation quality. As translation for every set of parameters with a large training set is computationally expensive, a small training corpus of size 15k sentence-pairs was chosen to tune the parameters. As alignment is also expensive, only four iterations of IBM Model 1 and two iterations of HMM alignments were performed. The lexical translation table obtained from the alignment of 15k sentence-pairs showed that 798 source words had more than two possible translations. The best N was found to be 42 which contained function words and content words.

Top ranking candidates based on each of the features, $f_1(w_j)$, $f_2(w_j)$, $f_3(w_j)$ and $f_4(w_j)$ along with their tuned combination (eqn. 1) are given in Table 2. Examples of target constraints extracted for the words, “*a*”, “*the*” and “*india*”, are given in Table 1. The target constraints of “*the*” mostly include case-markers. Blind removal of all instances of “*the*” in the training corpus resulted in a drop of 0.3 BLEU points over the constraint-

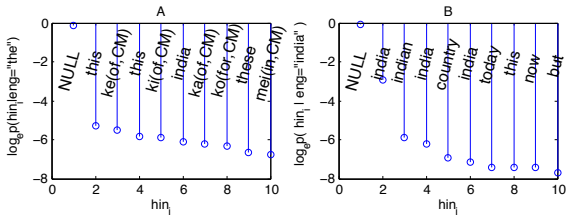

Figure 3: Lexical translation probabilities for source words, “*the*” and “*india*” obtained without removal of redundant source words from the training corpus. CM: case-markers in Hindi.

Score	Base	Rem
3-gram BLEU	12.09	18.57
4-gram BLEU	7.06	12.35
f-score (4k)	49.08	50.74

Table 3: Alignment scores (w.r.t 4k hand-aligned sentence-pairs), n-gram BLEU scores and modified precision scores (p_k) obtained with (Rem) and without (Base) removal of redundant words.

based removal of “*the*”. The reason is that, the case-markers in Hindi have no correspondences in English and require spurious words on the source for their generation. This suggests that redundancy has to be tackled both in the source and in the target to bring about balance in the number of words in the source and target sentences.

A small set of 4k sentence-pairs (from the 15k sentence-pairs) were hand-aligned to compute the f-score on the automatic alignments. While computing the alignment score, alignments of redundant source words were not considered in order to see the alignment improvements of other source words in the training corpus. Improvements in alignment and translation scores w.r.t the baseline are shown in Table 3. Statistically significant improvement ($p<0.001$) of 5.29 BLEU points in translation quality was seen on the test set with removal of redundant words.

5 Conclusion and Future Work

A novel approach to detect and remove redundancy was proposed which gave improvements in alignment as well as translation quality.

Future work will concentrate on removing redundancy even in the target before aligning the corpus and during testing, perform post-processing to insert target words in the translation. It would be interesting to see the performance of the method with other language-pairs.

References

Tagyoung Chung and Daniel Gildea 2010. Effects of empty categories on machine translation, *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, pp. 636-645.

Christopher Dyer, Smaranda Muresan and Philip Resnik. 2008. Generalizing Word Lattice Translation, *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pp. 1012-1020.

Gumwon Hong, Seung-Wook Lee and Hae-Chang Rim. 2009. Bridging morpho-syntactic gap between source and target sentences for English-Korean statistical machine translation, *In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (ACLShort '09)*, pp. 233-236.

Philipp Koehn, Franz Josef Och and Daniel Marcu, 2003. Statistical phrase-based translation, *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL '03) - Volume 1*, pp. 48-54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Caliison-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 ACL on Interactive Poster and Demonstration Sessions (ACL'07)*.

Jonghoon Lee, Donghyeon Lee and Gary Geunbae Lee. 2006. Improving Phrase-based Korean-English Statistical Machine Translation, *In Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech-ICSLP '06)*.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation, *In Proceedings of Human Language Technology conference-North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL-Short '04)*, pp. 57-60.

Chi-Ho Li, Dongdong Zhang, Mu Li, Ming Zhou and Hailei Zhang. 2008. An empirical study in source word deletion for phrase-based statistical machine translation, *In Proceedings of the Third Workshop on Statistical Machine Translation (StatMT '08)*, pp. 1-8.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL '03)*, Volume 1, pp. 160-167.

Kishore Papineni, 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 on Association for Computational Linguistics (ACL '02)*, pp. 311-318.

Michael James David Powell. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives, *Computer Journal*, pp. 155-162.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods, *Biometrics*, 1 ,80-83. *tool:http://faculty.vassar.edu/lowry/wilcoxon.html*.

Yuqi Zhang, Evgeny Matusov and Hermann Ney. 2009. Are Unaligned Words Important for Machine Translation ?, *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT '09)*, pp. 226-233.