Improving Word Alignment by Exploiting Adapted Word Similarity

Septina Dian Larasati
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
SIA TILDE
Riga, Latvia
larasati@ufal.mff.cuni.cz, septina@tilde.lv

Abstract

This paper presents a method to improve a word alignment model in a phrase-based Statistical Machine Translation system for a low-resourced language using a string similarity approach. Our method captures similar words that can be seen as semi-monolingual across languages, such as numbers, named entities, and adapted/loan words. We use several string similarity metrics to measure the monolinguality of the words, such as Longest Common Subsequence Ratio (LCSR), Minimum Edit Distance Ratio (MEDR), and we also use a modified BLEU Score (modBLEU).

Our approach is to add intersecting alignment points for word pairs that are orthographically similar, before applying a word alignment heuristic, to generate a better word alignment.

We demonstrate this approach on Indonesian-to-English translation task, where the languages share many similar words that are poorly aligned given a limited training data. This approach gives a statistically significant improvement by up to 0.66 in terms of BLEU score.

1 Introduction

Low-resourced languages do not have the luxury of having sufficient data to make a good statistical model. In some cases, those languages also do not have any additional language tools to make a linguistically motivated model. This limits the possibilities for low-resourced languages to gain a better translation quality in a Statistical Machine Translation (SMT) experiment.

Word alignment as the basic foundation in phrase-based SMT has gained significant attention in the research community. One of the most commonly applied word alignment approaches in a phrase-based SMT is to combine sets of alignment points learned from two directions (source-to-target and target-to-source). Another approach is to combine different sets of alignment points generated based on different motivations, such as linguistics and heuristics (Xiang et al., 2010). There are also work on using linguistics clues such as string similarity to harvest better word alignments (Bergsma and Kondrak, 2007) or by combining word-level and character-level models in SMT (Nakov and Tiedemann, 2012).

In this paper, we define an algorithm that adds intersecting alignment points on sets of alignment points learned from two different directions. Those added points are points between two similar words (measured by a string similarity metric). Then we apply one of the commonly used word alignment heuristics, MOSES's `grow-diag-final` (*gdfa*), on the new sets of alignment points to generate a better word alignment.

2 The Language Pair

In this work, we choose Indonesian as the low-resourced language and pair it with English. Indonesian-English SMT research is not so prolific. Similar work was done by (Nakov and Ng, 2009) for translating a resource-poor language, Indonesian, to English by using Malay as a pivot language. But most of the related SMT research is done for Malay,

a mutually intelligible language to Indonesian.

Because of the Indonesian complex morphology and the limited data availability, pairing Indonesian and English in an SMT experiment raises a challenge on creating a good word alignment model. Here we try to exploit their orthographically similar word pairs to improve the word alignment.

Some languages that are highly influenced by other languages tend to have similar words. Some of the words may be slightly different in their modified forms. In some cases, we intuitively know how to align words across languages by simply observing their word form.

Although Indonesian has a complex morphology, such as affixation and even reduplication, several Indonesian new words are highly influenced by English, and Indonesian tends to have some loan or adapted words. The words' orthographic similarity can be easily measured, since both languages have the same alphabet. Here we list some word pair examples that we consider semi-monolingual since they are orthographically similar.

Named Entity - Some named entities are poorly aligned because they are scarce in a given limited data. Those named entities both in Indonesian and English have a similar form and can be detected easily, even in their affixed forms, e.g. Indonesian '*Blackberryku*' and the corresponding English 'my Blackberry'.

Loan and Adapted Words - Indonesian adapts several English words and morphemes, e.g.

en	$\leftrightarrow$	***id***
distribution	$\leftrightarrow$	*distribusi*
idealist	$\leftrightarrow$	*idealis*
industry	$\leftrightarrow$	*industri*
department	$\leftrightarrow$	*departemen*
computer	$\leftrightarrow$	*komputer*
president	$\leftrightarrow$	*presiden*

Number and Radix Point - Numbers can come in different combination and are often scarce. They are easy to detect although Indonesian and English radix point are different, where Indonesian uses the comma symbol to separate the integer from the fraction while English uses the dot symbol, e.g. a thousand is 1.000,0 in Indonesian and 1,000.0 in English.

3 Improving the Word Alignment

We improve the word alignment by adding alignment points in the source-to-target (*f2e*) and/or the target-to-source (*e2f*) alignment to add more intersecting alignment points among them. Those intersecting alignment points are added on the word pairs that we consider similar. Then we apply a word alignment heuristic on the new *f2e* and *e2f* sets that now have more intersecting alignment points, to make a new word alignment.

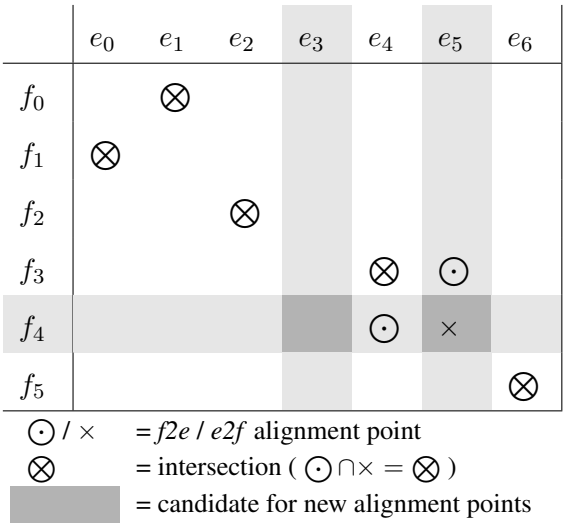

Figure 1: Choosing the candidates for the word pairing and the candidate positions for the new intersecting alignment points.

Suppose $f2e_{ij}$ is a source-to-target alignment link between the *i*-th source word (f_i) and the *j*-th target word (e_j) and $e2f_{ij}$ is a target-to-source alignment between f_i and e_j. Our approach to improve the word alignment is as follow:

1. We choose the source candidate words (*c(f)*) and the target candidate words (*c(e)*), where they are words that are not included in any intersecting alignment points, as illustrated in Figure 1.

2. We pair each *c(f)* to each *c(e)* and score their *string similarity* (*ss*).

3. We choose which pair to be aligned using our *filtering method*.

4. We add alignment points in the *f2e* and/or *e2f* alignment so that the alignment points for the chosen word pair intersect.

5. We apply the `grow-diag-final` (*gdfa*) heuristic[1] on the new *f2e* and *e2f* alignment to produce the new word alignment.

3.1 String Similarity Score

In this work, we uses three different string similarity measures, namely Longest Common Subsequence Ratio (LCSR), Minimum Edit Distance Ratio (MEDR), and a modified BLEU Score (modBLEU). We uses the three metrics to measure our string similarity score (*ss*). Here, we compare the modified BLEU formula to commonly known string similarity metrics, LCSR and MEDR. The LSCR and MEDR formula can be found in Figure 2.

$$ss(f_i, e_j) = LCSR(f_i, e_j) = \frac{|LCS(f_i,e_j)|}{max(|f_i|,|e_j|)} \quad \text{(a)}$$

$$ss(f_i, e_j) = MEDR(f_i, e_j) = 1 - \frac{|MED(f_i,e_j)|}{max(|f_i|,|e_j|)} \quad \text{(b)}$$

Figure 2: The Longest Common Subsequence Ratio (LCSR) and the Minimum Edit Distance Ratio (MEDR) formula for the string similarity metric.

In the modified BLEU, we split the words into characters and we use a modified BLEU on the character level as our string similarity score to measure the characters n-gram precision between the two words.

We score the word pairs using the modified BLEU in two directions: the source word as the hypothesis and the target word as the reference then vice versa. Then we average the two scores. Instead of using at most 4-grams counts in the original BLEU score formula, we modified the formula so that it also consider words with length less than four characters. Below is the formula for the modified BLEU given the length of the hypothesis (*c*) and the length of the reference (*r*).

3.2 Filtering Method

We set an *ss* score threshold to filter the word pairs. We only consider word pairs with a score equal to the threshold and above. We sort the candidate pairs by their *ss* score then by their source token order (*i*) and target token order (*j*). In this way, we pick the most similar word pairs first and then word pairs that occur earlier in the sentence. We assume that the similar words have the same order of occurrence in the sentence.

$$BLEU_m(f_i, e_j) = BP \bullet exp(\textstyle\sum_{n=1}^{chk} log(p_n)) \quad (1)$$

$$BP = min(1, e^{1-r/c}) \quad (2)$$

$$chk = min(4, r, c) \quad (3)$$

$$ss = (BLEU_m(f_i, e_j) + BLEU_m(e_j, f_i))/2 \quad (4)$$

Figure 3: The modified BLEU formula for the string similarity metric.

All the newly added alignment points have to be one-to-one aligned. We discard any new alignment point that violates this condition, as we pick the word pairs.

Consider Figure 1, if f_4, e_3, and e_5 all are the same word '*street*', so that $BLEU_m(f_4, e_3) = BLEU_m(f_4, e_5) = 100\%$, only a link between f_4 and e_3 is added, because the pair occurs earlier in the sentence and adding another link between f_4 and e_5 will violates the one-to-one alignment. If f_4, e_3, and e_5 are '*street*', '*streen*', '*street*' respectively, only a link between f_4 and e_5 is added, because it is chosen first for its better score.

4 Experiment

4.1 Data

The corpus we use in this work is the IDENTIC (Larasati, 2012) Indonesian-English parallel corpus. Our training, tuning, and testing data contain around 43K, 1K, and 1K parallel sentences respectively. The sentences are taken randomly without replacement from the corpus.

4.2 Common Setting

The SMT system is in lowercased-to-lowercased Indonesian-to-English translation direction. We use the state-of-the-art phrase-based SMT system MOSES (Koehn et al., 2007). We use GIZA++ tool

[1] http://www.statmt.org/moses/?n=FactoredTraining.AlignWords

(Och and Ney, 2003) to build the bidirectional sets of alignment points (*f2e* and *e2f*).

For the *baseline* system, we run the MOSES *gdfa* heuristic on the initial *f2e* and *e2f*. And for the experiment systems, we apply our algorithm to the initial *f2e* and *e2f* to generate the new sets and then we apply the same *gdfa* heuristic on the new sets. This makes the *gdfa* heuristic algorithm starts with more intersecting alignment points.

We create the English Language Model (LM) using SRILM (Stolcke, 2002) on the English Europarl corpus. The quality of the translation results are measured using BLEU score (Papineni et al., 2002) and pairwise bootstrapping significance test (Koehn, 2004).

4.3 Result

We set up the *baseline* system and the *exact* system. Then we created five experimental SMT systems that are set with different *ss* score thresholds, namely 90, 80, 70, 60, and 50.

The *exact* system aligns word pairs that are orthographically equal. Here the algorithm successfully aligns the foreign words ('supreme', 'court', etc), named entities ('wall' 'street', 'telkom', 'jakarta'), numbers ('1.28', '4.1', etc), and punctuations.

As we use different thresholds, the algorithm can pair Indonesian affixed words ('*uraniumnya*' with 'uranium'), adapted words ('*internasional*' with 'international' or '*kwartet*' and 'quartet'), and different number formatting ('0,85863' with '0.85863'). It also captures and pairs some inconsistent number formatting such as '6.5' and '6.50' and some misspelling such as '*streen*' and 'street' of the word 'wall street'.

In general, the translation quality increases when we add links for very similar words measured by any of the string similarity metrics. When we use the modified BLEU metric, the system's BLEU score is increasing in a logarithmic scale when the threshold is between 60 and 100.

But as the threshold set to a lower value, it wrongly aligns some short stopwords such as the Indonesian '*ini*' (this) with the English preposition 'in' and the Indonesian '*itu*' (that) with the English personal pronoun 'it', which makes the translation quality become poor. When we use the LCSR and MEDR metric, the translation quality decreases earlier with a bigger threshold.

	System	Δ to *exact*	BLEU
	baseline	-9469	27.25
	exact	0	*27.62
modBLEU	th*ss*-90	16	**27.83
	th*ss*-80	390	**27.87
	th*ss*-70	857	**27.89
	th*ss*-60	1457	**27.91
	th*ss*-50	3543	27.23
LCSR	th*ss*-90	107	**27.82
	th*ss*-80	1321	*27.52
	th*ss*-70	2813	27.42
	th*ss*-60	8112	27.12
	th*ss*-50	30214	*27.40
MEDR	th*ss*-90	83	*27.74
	th*ss*-80	975	*27.70
	th*ss*-70	2185	**28.03
	th*ss*-60	5693	27.25
	th*ss*-50	18037	27.12

* / **) 90% / 95% statistically significant

Table 1: SMT systems evaluation in term of BLEU score. The experiment systems with different thresholds are named th*ss*-[*threshold*]. Δ to *exact* is the number of the added intersection points compared to the *exact* system.

Table 1 summarizes the number of the added intersecting alignment points and the evaluation for the *baseline* and the experiment systems.

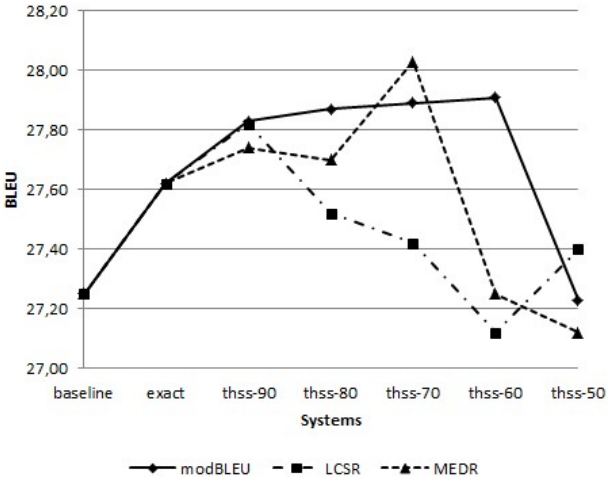

Figure 4: The *baseline* and the experimental SMT systems translation quality in terms of BLEU Score.

5 Conclusion

Our method captured similar words that are semi-monolingual across languages, such as numbers, named entities, and adapted words. We used this information as a clue to add alignment points. We showed that adding good quality intersecting alignment points before applying the *gdfa* heuristic helps to gain a better translation quality for a Indonesian-to-English SMT system. We used LCSR and MEDR as string similarity metrics and we also introduced another metric, a modified BLEU formula. We still found some word pairs that are wrongly aligned and most of them are stopwords. Modifying the string similarity formula or the filtering method so that it does not capture these stopwords will be a good future improvement.

Acknowledgments

The research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA), by the grant LC536 Centrum Komputační Lingvistiky of the Czech Ministry of Education, and this work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

Shane Bergsma and Grzegorz Kondrak. 2007. Alignment-based discriminative string similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 656–663, Prague, Czech Republic, June. Association for Computational Linguistics.

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-Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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 Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Septina Dian Larasati. 2012. IDENTIC corpus: Morphologically enriched indonesian-english parallel corpus. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1367, Singapore, August. Association for Computational Linguistics.

Preslav Nakov and Hwee Tou Ng. 2011. Translating from morphologically complex languages: a paraphrase-based approach. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL2011), Portland, Oregon, USA*.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305, Jeju Island, Korea, July. 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.

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*, pages 311–318. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*.

Bing Xiang, Yonggang Deng, and Bowen Zhou. 2010. Diversify and combine: Improving word alignment for machine translation on low-resource languages. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 22–26, Uppsala, Sweden, July. Association for Computational Linguistics.