英日機械翻訳における語順評価の有効性

賀沢 秀人, David Talbot , 妹尾 政和, Jason Katz-Brown, 市川 宙, Franz Och
Google, Inc.
{kazawa, talbot, seno, jasonkb, ichikawa, och}@google.com

1 Word Order Metrics in MT

Automatic evaluation is important for NLP in general. Then there are many researches about automatic evaluation in MT. Most of them measures n-gram matches between reference and system translations. But some authors pointed out that such metrics don't count the correctness of word order enough and auto-evaluation can be improved by incorporating word order information.[1, 2, 3]

Isozaki et al. proposed a word-order evaluation metric based on rank statistics, which regards translation system's output as word permutation of the reference translation.[1] They showed experimentally that the word order metrics are more predictive of the human evaluation of translation quality than the standard n-gram-based metric BLEU. However their approach uses unigram and bigram matches to align words between reference and system translations, which may end up with many words unaligned. Though they provided a workaround, brevity penalty which discounts overestimation caused by few matching n-grams, it doesn't allow to fully utilize word order information in the reference translation.

Birch et al. proposed a similar word metric based on word alignment and rank statistics.[2, 3] Their approach differs from Isozaki et al.'s that it works with any word alignment algorithm. With this flexibility, they argued that if metric's accuracy is very important, one can manually align words of source and reference and then automatically align words of reference and system translation through source-translation word alignment, which is often obtained as a by-product of the decoder. Then the word order information in reference can be fully utilized.

They conducted a validation experiment in [2]. They, however, used "fake" translations which were obtained by randomly reordering words of reference. Such artificial data can overestimate the accuracy of word order metric. Suppose that we translate English sentence "I was sad because he stole it." to Japanese and get the following translations. (Aligned English words are written in parenthesis in the order of Japanese words.)

T1 彼がそれを盗んだので私は悲しかった (he it stole because I sad was)

T2 私は悲しかったなぜなら彼がそれを盗んだから。 (I sad was because he it stole)

Both are arguably acceptable. But if we used lexicons in T1 with T2's word order, the translation would be much worse.

T2' 私は悲しかったので彼がそれを盗んだ

Birch et al.'s experiment in [2] would present T1 and T2', not T2, to human evaluators. Such comparison would result in favor of T1. In general, artificial translations composed from lexical choice and word order of different systems can be much worse than original translations.

In this paper, we present experimentally that word order of real system translation is strongly correlated to human judgement of its quality, in other words, Birch et al.'s argument is correct. Additionally we propose a novel way of creating word alignment manually where human annotators create translations and word alignment data simultaneously, which lead to densely word aligned translations.

2 Translation with Alignment

2.1 Alignment-oriented Translation

Usually word alignment annotation is done separately from translation. With this approach, however, alignment can be difficult when translations are quite different from "literal" translations, which is often the case for fluent translations.

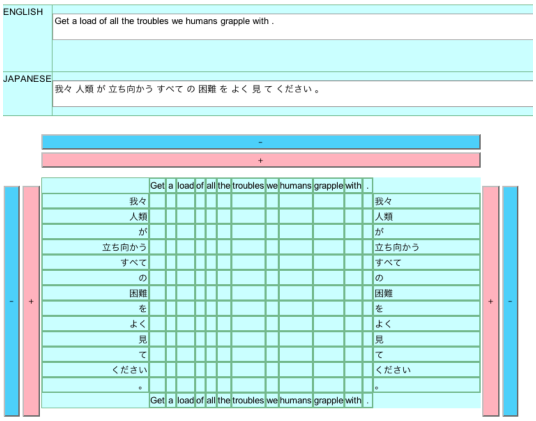
For example, English sentence "The scene made me happy." could be translated to "私はそのシーンを見て幸せな気分になった。". Many annotators would agree that "The scene" is aligned to "そのシーンを" and "me" to "私は". But there seems no obvious alignment for the remaining words. This kind of difficulties result in inconsistent annotations and inefficiency of the annotation job.

To avoid this problem, we can take advantage of the fact that there are usually many good/acceptable translations, and it is likely that some of them are easier to align words. For the above example, ”そのシーンは私を幸せにした。” would be such translation.

On the other hand, such alignment-oriented translations may sacrifice quality. The second translation ”そのシーンは私を幸せにした。” is arguably less natural than the first translation ”私はそのシーンを見て幸せな気分になった。”. So there is a trade-off between the ease of alignment and the quality of translation. We will revisit issues later.

2.2 Translation-with-Alignment Tool

To create consistent word alignment data, we developed a browser-based GUI which allows human annotators to create and modify translation and word alignment simultaneously. (Fig. 1)



ENGLISH

Get a load of all the troubles we humans grapple with .

JAPANESE

我々 人類 が 立ち向かう すべて の 困難 を よく 見 て ください 。

	Get	a	load	of	all	the	troubles	we	humans	grapple	with	.	
我々													我々
人類													人類
が													が
立ち向かう													立ち向かう
すべて													すべて
の													の
困難													困難
を													を
よく													よく
見													見
て													て
ください													ください
。													。
	Get	a	load	of	all	the	troubles	we	humans	grapple	with	.	

Figure 1: Screenshot of the translation-with-alignment tool

A source sentence is presented to an annotator in the top box. The annotator fills a draft translation in the second box, which is automatically segmented into words and filled into the alignment matrix below. Then the annotator chooses word alignment by clicking the matrix cell which is the intersection of the corresponding word column and row. To encourage annotators to modify translations so that they can align words as many as possible, the GUI allows to group consecutive words into a chunk and to modify translation wordings of such chunk in situ.

Annotators were given a couple of guidelines to make word alignment consistent.

- The translation should be a sequence of word chunks, each of which should be aligned to a chunk in the source sentence. We loosely define chunk as ”a unit in sentence such as word, phrase, clause and bunsetsu.”
- The above rule should be applied recursively i.e, if some chunk consists of smaller chunks, each of such smaller chunks should be aligned to a subchunk in the corresponding chunk.
- Slight unnaturalness in translation is acceptable if it is difficult to follow the above rules without making translation unnatural.

3 Experiments

3.1 Word Order Metrics

In this paper, we measure the correctness of word order in two ways: Kendall's tau and Fuzzy Reordering Score (FRS). In the following descriptions, it is assumed that both reference and system translations are word-aligned to source sentence.

Fuzzy Reordering Score To compute Fuzzy Reordering Score, first, the sequence of matched words (”matches”) is divided into the fewest possible number of “chunks” such that the matched words in each chunk are adjacent (in both translations) and in identical word order. The number of chunks (ch) and the number of matches (m) is then used to calculate Fuzzy Reordering Score.

$$\mathrm{FRS} = 1 - \frac{ch - 1}{m - 1} \tag{1}$$

When the strings are empty or have just one word, we define the score is 1.0. Note that a similar metric is used by Lavie et al. when they introduced a word order penalty into METEOR.[4]

Kendall's tau Kendall's tau is used in the other authors' work on word order evaluation. Let r_i and t_i be the indices of aligned source word of i-th word of reference and system translations respectively. Any pair of (r_i, t_i) and (r_j, t_j) is said to be concordant if and only if $r_i < r_j$ and $t_i < t_j$, or $r_i > r_j$ and $t_i > t_j$. Other pairs are said to be discordant. With these, we define Kendall's tau (Tau) in this paper as follows.

$$\mathrm{Tau} = \frac{\#(\text{concordant pairs}) - \#(\text{discordant pairs})}{\#(\text{all pairs})} \tag{2}$$

3.2 Correlation of Word Order and Translation Quality

System-level correlation We created 9 phrase-based statistical machine translation systems to study system-level correlations. All systems do translations in three stages. First it parses the source sentence into a dependency tree. Then it reorders source words so that the result word sequence looks like a sentence in the target language if translated word-by-word. Finally the reordered word sequence is translated monotonically with a phrase table and a language model, which were trained with a large collection of web data.

Each experimental system has a different set up in the parsing and in the reordering modules. We had five parsers, which were different in the size of training data and the parsing strategy, and five sets of reordering rules which were different in the complexity. The simplest rule set actually does nothing and the most complicated one determines the order of dependent words based on the category of head word. By combining different parsers and reordering rules, we created 9 systems for experiments.

We translated 500 English sentences randomly sampled from the Web by these systems and human raters evaluated each translation in a 7 point scale. The translations of the same source were presented at once to each rater. Other human raters created translations and word alignment data for the same set of English sentences with the tool mentioned above. We computed system-level averages of score by human, Fuzzy Reordering Score, Kendall's tau and BLEU. (Table 1)

You can see the two word order metrics, FRS and Tau, are more correlated to Human than BLEU. Especially BLEU is not predictive of the performance differences among System 4–9 where FRS and Tau are more predictive. There is one peculiarity that FRS and Tau show statistically significant difference between System 4 and 5 but Human says there is no significant difference.

Sentence-level correlation In the above experiment, we found that the inter-rater agreement of Human was low but the difference of Human scores were consistent. So we examined correlation of the differences of metric and human score. (Table 2) Since score differences are not independent variables, we randomly sampled 300 system-sentence pairs, which is 1.7% of all pairs, to avoid artificial biases.

In addition to those metrics, we also tried simple combinations. FRS and Tau are based on different views of measuring order so we also studied the correlation of their combination, here the sum of them. Also the previous works reported that the combination of word order metric and precision metric such as BLEU improved correlation. So we included two simple combination, the sum of FRS and BLEU, and a (weighted) sum of FRS, Tau and BLEU. Those results are also in Table 2.

Metric	Correlation
FRS	0.508
Tau	0.505
BLEU	0.409
FRS+Tau	0.546
FRS+BLEU	0.560
(FRS+Tau)/2 + BLEU	0.588

Table 2: Sentence-level correlation of evaluation metrics to human judgement score

Human Score	W/ Align	W/O Align
6	18.5%	35.5%
5	25.0%	27.5%
4	31.0%	20.0%
3	13.5%	8.5%
2	7.0%	6.0%
1	2.0%	2.0%
0	3.0%	0.5%

Table 3: Distribution of subjective scores of with/without alignment translations

Among non-combined metrics, the correlations are relatively high for FRS and Tau. As many authors pointed out, the result confirmed that BLEU is not so well correlated to Human at sentence level. On the other hand, we see improvements with the simple combination of the metrics. This seems to suggest that more improvement could be seen by optimizing weights of combination and/or adding more metrics.

3.3 Quality of With-alignment Translations

The quality of with-alignment translations may be worse than normal translations as we mentioned. And it is not reasonable to rely on word orders which don't produce good translations. We compared 200 with/without-alignment translations of random English Web sentences. The raters were presented both types of translations at once and asked to give a 7-point score to each translations. (Table 3)

We looked at with/without-alignment translation pairs whose with-alignment score is less than 5 but without-alignment score is 6. There are 44 such translation pairs. We manually checked with-alignment translations to see if one could create good translations by replacing lexical choices while keeping the word order. We found that 37 translations would be as good as without-alignment translations if lexical choices were better. We found the other 7

System	Human	FRS	Tau	BLEU
1	1.24 [1.15,1.34]	0.392 [0.375,0.410]	0.254 [0.215,0.291]	0.061 [0.051,0.072]
2	1.83 [1.71,1.95]	0.393 [0.374,0.412]	0.501 [0.466,0.536]	0.083 [0.071,0.096]
3	1.85 [1.73,1.98]	0.541 [0.521,0.561]	0.487 [0.456,0.517]	0.091 [0.078,0.105]
4	2.17 [2.04,2.30]	0.724 [0.704,0.744]	0.711 [0.681,0.739]	0.110 [0.096,0.124]
5	2.19 [2.06,2.33]	0.618 [0.599,0.637]	0.654 [0.628,0.681]	0.110 [0.096,0.125]
6	2.31 [2.18,2.44]	0.747 [0.728,0.767]	0.749 [0.721,0.776]	0.113 [0.098,0.128]
7	2.40 [2.27,2.53]	0.759 [0.740,0.777]	0.756 [0.729,0.783]	0.114 [0.099,0.129]
8	2.44 [2.31,2.57]	0.775 [0.756,0.794]	0.789 [0.764,0.813]	0.118 [0.103,0.133]
9	2.50 [2.37,2.63]	0.779 [0.760,0.798]	0.788 [0.761,0.813]	0.119 [0.104,0.134]

Table 1: System-level average of human judge score (Human), Fuzzy Reordering Score (FRS), Kendall's tau (Tau) and BLEU. Each record shows the average score and the 95% confidence interval obtained by bootstrap.

translations had actually wrong word order.

We saw two major patterns in the bad translations. First some translations were too verbose.

Source

> Persistence is the quality that we need to obtain our goals.

With-alignment translation

> 粘り強さは私達が私達の目標を達成するのに必要であるという特性です。

Without-alignment translation

> 粘り強さは、目標達成に必要な資質です。

The with-alignment translation has two ”私達”, which are the translations of ”we” and ”our”. Both refer to general public which is usually untranslated in Japanese. The alignment requirement, however, doesn't allow not to translate ”we” and ”our”. One can argue that this kind of abbreviation involve context processing which is beyond the scope of sentence-by-sentence translation. We could argue that these words should be kept translated and be removed in other modules if necessary.

The second major pattern is bad lexical choice.

Source

> To help you make your decision, I have tested and reviewed the best-rated eBook compilers currently available.

With-alignment translation

> あなたがあなたの決断を下すのを手助するために、私は現在入手可能な最高評価の電子書籍のコンパイラーをテストしそして見直しました。

Without-alignment translation

> 皆様の意思決定をお手伝いするために、現在入手可能な中で最高の評価を得ている電子ブック編集機をテストおよびレビューしました。

In addition to verbose ”あなた”, the with-alignment translation has less fluent or wrong translations ”コンパイラー” and ”見直しました”. You can see, however, that these bad words don't affect the relevancy of word order. Actually the word order of the without-alignment translation is exactly the same.

4 Conclusions

We have presented that a novel way to create word-alignment data, which allows us to align words consistently. Our experiments showed that the word order metrics using the word-alignment data are correlated to human evaluation of translation better than BLEU at system level. Also the metrics showed modest correlation at sentence level.

References

[1] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh and Hajime Tsukada. *Automatic Evaluation of Translation Quality for Distant Language Pairs*, Proc. of EMNLP-2010, pp.944–952.

[2] Alexandra Birch, Miles Osborne and Phil Blunsom. *Metrics for MT evaluation: evaluating reordering*, Machine Translation, 24(1) pp.15–26.

[3] Aleandra Birch and Miles Osborne. *LRscore for evaluating lexical and reordering quality in MT*, Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp.327–332.

[4] Alon Lavie and Michael Denkowski. *The METEOR Metric for Automatic Evaluation of Machine Translation*, Machine Translation, 23(2-3), pp.105–115