A Lightweight Evaluation Framework for Machine Translation Reordering

David Talbot[1] and **Hideto Kazawa**[2] and **Hiroshi Ichikawa**[2]
Jason Katz-Brown[2] and **Masakazu Seno**[2] and **Franz J. Och**[1]

[1] Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
{talbot, och}@google.com

[2] Google Japan
Roppongi Hills Mori Tower
6-10-1 Roppongi, Tokyo 106-6126
{kazawa, ichikawa}@google.com
{jasonkb, seno}@google.com

Abstract

Reordering is a major challenge for machine translation between distant languages. Recent work has shown that evaluation metrics that explicitly account for target language word order correlate better with human judgments of translation quality. Here we present a simple framework for evaluating word order independently of lexical choice by comparing the system's reordering of a source sentence to reference reordering data generated from manually word-aligned translations. When used to evaluate a system that performs reordering as a preprocessing step our framework allows the parser and reordering rules to be evaluated extremely quickly without time-consuming end-to-end machine translation experiments. A novelty of our approach is that the translations used to generate the reordering reference data are generated in an *alignment-oriented* fashion. We show that how the alignments are generated can significantly effect the robustness of the evaluation. We also outline some ways in which this framework has allowed our group to analyze reordering errors for English to Japanese machine translation.

1 Introduction

Statistical machine translation systems can perform poorly on distant language pairs such as English and Japanese. Reordering errors are a major source of poor or misleading translations in such systems (Isozaki et al., 2010). Unfortunately the standard evaluation metrics used by the statistical machine translation community are relatively insensitive to the long-distance reordering phenomena encountered when translating between such languages (Birch et al., 2010).

The ability to rapidly evaluate the impact of changes on a system can significantly accelerate the experimental cycle. In a large statistical machine translation system, we should ideally be able to experiment with separate components without retraining the complete system. Measures such as perplexity have been successfully used to evaluate language models independently in speech recognition eliminating some of the need for end-to-end speech recognition experiments. In machine translation, alignment error rate has been used with some mixed success to evaluate word-alignment algorithms but no standard evaluation frameworks exist for other components of a machine translation system (Fraser and Marcu, 2007).

Unfortunately, BLEU (Papineni et al., 2001) and other metrics that work with the final output of a machine translation system are both insensitive to reordering phenomena and relatively time-consuming to compute: changes to the system may require the realignment of the parallel training data, extraction of phrasal statistics and translation of a test set. As training sets grow in size, the cost of end-to-end experimentation can become significant. However, it is not clear that measurements made on any single part of the system will correlate well with human judgments of the translation quality of the whole system.

Following Collins et al. (2005a) and Wang (2007), Xu et al. (2009) showed that when translating from English to Japanese (and to other SOV languages such as Korean and Turkish) applying reordering as

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 12–21,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

a preprocessing step that manipulates a source sentence parse tree can significantly outperform state-of-the-art phrase-based and hierarchical machine translation systems. This result is corroborated by Birch et al. (2009) whose results suggest that both phrase-based and hierarchical translation systems fail to capture long-distance reordering phenomena.

In this paper we describe a lightweight framework for measuring the quality of the reordering components in a machine translation system. While our framework can be applied to any translation system in which it is possible to derive a token-level alignment from the input source tokens to the output target tokens, it is of particular practical interest when applied to a system that performs reordering as a preprocessing step (Xia and McCord, 2004). In this case, as we show, it allows for extremely rapid and sensitive analysis of changes to parser, reordering rules and other reordering components.

In our framework we evaluate the reordering proposed by a system separately from its choice of target words by comparing it to a reference reordering of the sentence generated from a manually word-aligned translation. Unlike previous work (Isozaki et al., 2010), our approach does not rely on the system's output matching the reference translation lexically. This makes the evaluation more robust as there may be many ways to render a source phrase in the target language and we would not wish to penalize one that simply happens not to match the reference.

In the next section we review related work on reordering for translation between distant language pairs and automatic approaches to evaluating reordering in machine translation. We then describe our evaluation framework including certain important details of how our reference reorderings were created. We evaluate the framework by analyzing how robustly it is able to predict improvements in subjective translation quality for an English to Japanese machine translation system. Finally, we describe ways in which the framework has facilitated development of the reordering components in our system.

2 Related Work

2.1 Evaluating Reordering

The ability to automatically evaluate machine translation output has driven progress in statistical machine translation; however, shortcomings of the dominant metric, BLEU (Papineni et al., 2001) , particularly with respect to reordering, have long been recognized (Callison-burch and Osborne, 2006). Reordering has also been identified as a major factor in determining the difficulty of statistical machine translation between two languages (Birch et al., 2008) hence BLEU scores may be most unreliable precisely for those language pairs for which statistical machine translation is most difficult (Isozaki et al., 2010).

There have been many results showing that metrics that account for reordering are better correlated with human judgements of translation quality (Lavie and Denkowski, 2009; Birch and Osborne, 2010; Isozaki et al., 2010). Examples given in Isozaki et al. (2010) where object and subject arguments are reversed in a Japanese to English statistical machine translation system demonstrate how damaging reordering errors can be and it should therefore not come as a surprise that word order is a strong predictor of translation quality; however, there are other advantages to be gained by focusing on this specific aspect of the translation process in isolation.

One problem for all automatic evaluation metrics is that multiple equally good translations can be constructed for most input sentences and typically our reference data will contain only a small fraction of these. Equally good translations for a sentence may differ both in terms of lexical choice and word order. One of the potential advantages of designing a metric that looks only at word order, is that it may, to some extent, factor out variability along the dimension of the lexical choice. Previous work on automatic evaluation metrics that focus on reordering, however, has not fully exploited this.

The evaluation metrics proposed in Isozaki et al. (2010) compute a reordering score by comparing the ordering of unigrams and bigrams that appear in both the system's translation and the reference. These scores are therefore liable to overestimate the reordering quality of sentences that were poorly translated. While Isozaki et al. (2010) does propose a work-around to this problem which combines the reordering score with a lexical precision term, this clearly introduces a bias in the metric whereby poor translations are evaluated primarily on their lexical choice and good translations are evaluated more on the basis of their word order. In our experience word order is particularly poor in those sentences that have the lowest lexical overlap with reference translations; hence we would like to be able to compute the quality of reordering in all sentences independently of the quality of their lexical choice.

Birch and Osborne (2010) are closer to our approach in that they use word alignments to induce a permutation over the source sentence. They compare a source-side permutation generated from a word alignment of the reference translation with one generated from the system's using various permutation distances. However, Birch and Osborne (2010) only demonstrate that these metrics are correlated with human judgements of translation quality when combined with BLEU score and hence take lexical choice into account.

Birch et al. (2010) present the only results we are aware of that compute the correlation between human judgments of translation quality and a reordering-only metric independently of lexical choice. Unfortunately, the experimental set-up there is somewhat flawed. The authors 'undo' reorderings in their reference translations by permuting the reference tokens and presenting the permuted translations to human raters. While many machine translation systems (including our own) assume that reordering and translation can be factored into separate models, e.g. (Xia and McCord, 2004), and perform these two operations in separate steps, the latter conditioned on the former, Birch et al. (2010) are making a much stronger assumption when they perform these simulations: they are assuming that lexical choice and word order are entirely *independent*. It is easy to find cases where this assumption does not hold and we would in general be very surprised if a similar change in the reordering component in our system did not also result in a change in the lexical choice of the system; an effect which their experiments are unable to model.

Another minor difference between our evaluation framework and (Birch et al., 2010) is that we use a reordering score that is based on the minimum number of chunks into which the candidate and reference permutations can be concatenated similar to the reordering component of METEOR (Lavie and Denkowski, 2009). As we show, this is better correlated with human judgments of translation quality than Kendall's τ. This may be due to the fact that it counts the number of 'jumps' a human reader has to make in order to parse the system's order if they wish to read the tokens in the reference word order. Kendall's τ on the other hand penalizes every pair of words that are in the wrong order and hence has a quadratic (all-pairs) flavor which in turn might explain why Birch et al. (2010) found that the square-root of this quantity was a better predictor of translation quality.

2.2 Evaluation Reference Data

To create the word-aligned translations from which we generate our reference reordering data, we used a novel *alignment-oriented* translation method. The method (described in more detail below) seeks to generate reference reorderings that a machine translation system might reasonably be expected to achieve. Fox (2002) has analyzed the extent to which translations seen in a parallel corpus can be broken down into clean phrasal units: they found that most sentence pairs contain examples of reordering that violate phrasal cohesion, i.e. the corresponding words in the target language are not completely contiguous or solely aligned to the corresponding source phrase. These reordering phenomena are difficult for current statistical translation models to learn directly. We therefore deliberately chose to create reference data that avoids these phenomena as much as possible by having a single annotator generate both the translation and its word alignment. Our word-aligned translations are created with a bias towards simple phrasal reordering.

Our analysis of the correlation between reordering scores computed on reference data created from such alignment-oriented translations with scores computed on references generated from standard professional translations of the same sentences suggests that the alignment-oriented translations are more useful for evaluating a current state-of-the-art system. We note also that while prior work has conjectured that automatically generated alignments are a suitable replacement for manual alignments in the context of reordering evaluation (Birch et al., 2008), our results suggest that this is not the case at least for the language pair we consider, English-Japanese.

3 A Lightweight Reordering Evaluation

We now present our lightweight reordering evaluation framework; this consists of (1) a method for generating reference reordering data from manual word-alignments; and (2) a reordering metric for scoring a sytem's proposed reordering against this reference data; and (3) a stand-alone evaluation tool.

3.1 Generating Reference Reordering Data

We follow Birch and Osborne (2010) in using reference reordering data that consists of permuations of source sentences in a test set. We generate these from word alignments of the source sentences to reference translations. Unlike previous work, however, we have the same annotator generate both the reference translation and the word alignment. We also explicitly encourage the translators to generate translations that are easy to align even if this does result in occasionally unnatural translations. For instance in English to Japanese translation we require that all personal pronouns are translated; these are often omitted in natural Japanese. We insist that all but an extremely small set of words (articles and punctuation for English to Japanese) be aligned. We also disprefer non-contiguous alignments of a single source word and require that all target words be aligned to at least one source token. In Japanese this requires deciding how to align particles that mark syntactic roles; we choose to align these together with the content word (*jiritsu-go*) of the corresponding constituent (*bunsetsu*). Asking annotators to translate and perform word alignment on the same sentence in a single session does not necessarily increase the annotation burden over stand-alone word alignment since it encourages the creation of *alignment-friendly* translations which can be aligned more rapidly. Annotators need little special background or training for this task, as long as they can speak both the source and target languages.

To generate a permutation from word alignments we rank the source tokens by the position of the first target token to which they are aligned. If multiple source tokens are aligned to a single target word or span we ignore the ordering within these source spans; this is indicated by braces in Table 2. We place unaligned source words immediately before the next aligned source word or at the end of the sentence if there is none. Table 2 shows the reference reordering derived from various translations and word alignments.

3.2 Fuzzy Reordering Score

To evaluate the quality of a system's reordering against this reference data we use a simple *reordering metric* related to METEOR's reordering component (Lavie and Denkowski, 2009) . Given the reference permutation of the source sentence σ_{ref} and the system's reordering of the source sentence σ_{sys} either generated directly by a reordering component or inferred from the alignment between source and target phrases used in the decoder, we align each word in σ_{sys} to an instance of itself in σ_{ref} taking the first unmatched instance of the word if there is more than one. We then define C to be the number chunks of contiguously aligned words. If M is the number of words in the source sentence then the *fuzzy reordering score* is computed as,

$$\mathbf{FRS}(\sigma_{\mathrm{sys}}, \sigma_{\mathrm{ref}}) \quad = \quad 1 - \frac{C-1}{M-1}. \qquad (1)$$

This metric assigns a score between 0 and 1 where 1 indicates that the system's reordering is identical to the reference. C has an intuitive interpretation as the number of times a reader would need to jump in order to read the system's reordering of the sentence in the order proposed by the reference.

3.3 Evaluation Tool

While the framework we propose can be applied to any machine translation system in which a reordering of the source sentence can be inferred from the translation process, it has proven particularly useful applied to a system that performs reordering as a separate preprocessing step. Such *pre-ordering* approaches (Xia and McCord, 2004; Collins et al., 2005b) can be criticized for greedily committing to a single reordering early in the pipeline but in practice they have been shown to perform extremely well on language pairs that require long distance reordering and have been successfully combined with other more integrated reordering models (Xu et al., 2009).

The performance of a parser-based pre-ordering component is a function of the reordering rules and parser; it is therefore desirable that these can be evaluated efficiently. Both parser and reordering rules may be evaluated using end-to-end automatic metrics such as BLEU score or in human evaluations. Parsers may also be evaluated using intrinsic treebank metrics such as labeled accuracy. Unfortunately these metrics are either expensive to compute or, as we show, unpredictive of improvements in human perceptions of translation quality.

Having found that the fuzzy reordering score proposed here is well-correlated with changes in human judgements of translation quality, we established a stand-alone evaluation tool that takes a set of reordering rules and a parser and computes the reordering scores on a set of reference reorderings. This has become the most frequently used method for evaluating changes to the reordering component in our system and has allowed teams working on parsing, for instance, to contribute significant improvements quite independently.

4 Experimental Set-up

We wish to determine whether our evaluation framework can predict which changes to reordering components will result in statistically significant improvements in subjective translation quality of the end-to-end system. To that end we created a number of systems that differ only in terms of reordering components (parser and/or reordering rules). We then analyzed the corpus- and sentence-level correlation of our evaluation metric with judgements of human translation quality.

Previous work has compared either quite separate systems, e.g. (Isozaki et al., 2010), or systems that are artificially different from each other (Birch et al., 2010). There has also been a tendency to measure corpus-level correlation. We are more interested in comparing systems that differ in a realistic manner from one another as would typically be required in development. We also believe sentence-level correlation is more important than corpus-level correlation since good sentence-level correlation implies that a metric can be used for detailed analysis of a system and potentially to optimize it.

4.1 Systems

We carried out all our experiments using a state-of-the-art phrase-based statistical English-to-Japanese machine translation system (Och, 2003). During both training and testing, the system reorders source-language sentences in a preprocessing step using a set of rules written in the framework proposed by (Xu et al., 2009) that reorder an English dependency tree into target word order. During decoding, we set the reordering window to 4 words. In addition to the regular distance distortion model, we incorporate a maximum entropy based lexicalized phrase reordering model (Zens and Ney, 2006). For parallel training data, we use an in-house collection of parallel documents. These come from various sources with a substantial portion coming from the web after using simple heuristics to identify potential document pairs. We trained our system on about 300 million source words.

The reordering rules applied to the English dependency tree define a precedence order for the children of each head category (a coarse-grained part of speech). For example, a simplified version of the precedence order for child labels of a verbal head HEADVERB is: advcl, nsubj, prep, [other children], dobj, prt, aux, neg, HEADVERB, mark, ref, compl.

The dependency parser we use is an implementation of a transition-based dependency parser (Nivre, 2008). The parser is trained using the averaged perceptron algorithm with an early update strategy as described in Zhang and Clark (2008).

We created five systems using different parsers; here *targeted self-training* refers to a training procedure proposed by Katz-Brown et al. (2011) that uses our reordering metric and separate reference reordering data to pick parses for self-training: an n-best list of parses is generated for each English sentence for which we have reference reordering data and the parse tree that results in the highest fuzzy reordering score is added to our parser's training set. Parsers P3, P4 and P5 differ in how that framework is applied and how much data is used.

- P1 Penn Treebank, perceptron, greedy search
- P2 Penn Treebank, perceptron, beam search
- P3 Penn Treebank, perceptron, beam search, targeted self-training on web data

- P4 Penn Treebank, perceptron, beam search, targeted self-training on web data
- P5 Penn Treebank, perceptron, beam search, targeted self-training on web data, case insensitive

We also created five systems using the fifth parser (P5) but with different sets of reordering rules:

- R1 No reordering
- R2 Reverse reordering
- R3 Head final reordering with reverse reordering for words before the head
- R4 Head final reordering with reverse reordering for words after the head
- R5 Superset of rules from (Xu et al., 2009)

Reverse reordering places words in the reverse of the English order. Head final reordering moves the head of each dependency after all its children. Rules in R3 and R4 overlap significantly with the rules for noun and verb subtrees respectively in R5. Otherwise all systems were identical. The rules in R5 have been extensively hand-tuned while R1 and R2 are rather naive. System P5R5 was our best performing system at the time these experiments were conducted.

We refer to systems by a combination of parser and reordering rules set identifiers, for instance, system P2R5, uses parser P2 with reordering rules R5. We conducted two subjective evaluations in which bilingual human raters were asked to judge translations on a scale from 0 to 6 where 0 indicates nonsense and 6 is perfect. The first experiment (Parsers) contrasted systems with different parsers and the second (Rules) varied the reordering rules. In each case three bilingual evaluators were shown the source sentence and the translations produced by all five systems.

4.2 Meta-analysis

We perform a meta-analysis of the following metrics and the framework by computing correlations with the results of these subjective evaluations of translation quality:

1. Evaluation metrics: BLEU score on final translations, Kendall's τ and fuzzy reordering score on reference reordering data
2. Evaluation data: both manually-generated and automatically-generated word alignments on both standard professional and *alignment-oriented* translations of the test sentences

The automatic word alignments were generated using IBM Model 1 in order to avoid directional biases that higher-order models such as HMMs have.

Results presented in square parentheses are 95 percent confidence intervals estimated by bootstrap resampling on the test corpus (Koehn, 2004).

Our test set contains 500 sentences randomly sampled from the web. We have both professional and *alignment-friendly* translations for these sentences. We created reference reorderings for this data using the method described in Section 3.1. The lack of a broad domain and publically available Japanese test corpus makes the use of this nonstandard test set unfortunately unavoidable.

The human raters were presented with the source sentence, the human reference translation and the translations of the various systems simultaneously, blind and in a random order. Each rater was allowed to rate no more than 3 percent of the sentences and three ratings were elicited for each sentence. Ratings were a single number between 0 and 6 where 0 indicates nonsense and 6 indicates a perfectly grammatical translation of the source sentence.

5 Results

Table 2 shows four reference reorderings generated from various translations and word alignments. The automatic alignments are significantly sparser than the manual ones but in these examples the reference reorderings still seem reasonable. Note how the alignment-oriented translation includes a pronoun (translation for 'I') that is dropped in the slightly more natural standard translation to Japanese.

Table 1 shows the human judgements of translation quality for the 10 systems (note that P5R5 appears in both experiments but was scored differently as human judgments are affected by which other translations are present in an experiment). There is a clear ordering of the systems in each experiment and

1. Parsers	Subjective Score (0-6)	2. Rules	Subjective Score (0-6)
P1R5	2.173 [2.086, 2.260]	P5R1	1.258 [1.191, 1.325]
P2R5	2.320 [2.233, 2.407]	P5R2	1.825 [1.746, 1.905]
P3R5	2.410 [2.321, 2.499]	P5R3	1.849 [1.767, 1.931]
P4R5	2.453 [2.366, 2.541]	P5R4	2.205 [2.118, 2.293]
P5R5	2.501 [2.413, 2.587]	P5R5	2.529 [2.441, 2.619]

Table 1: Human judgements of translation quality for 1. Parsers and 2. Rules.

Metric	Sentence-level correlation	
	r	ρ
Fuzzy reordering	0.435	0.448
Kendall's τ	0.371	0.450
BLEU	0.279	0.302

Table 6: Pearson's correlation (r) and Spearman's rank correlation (ρ) with subjective translation quality at sentence-level.

we see that both the choice of parser and reordering rules clearly effects subjective translation quality.

We performed pairwise significance tests using bootstrap resampling for each pair of 'improved' systems in each experiment. Tables 3, 4 and 5 shows which pairs were judged to be statistically significant improvements at either 95 or 90 percent level under the different metrics. These tests were computed on the same 500 sentences. All pairs but one are judged to be statistically significant improvements in subjective translation quality. Significance tests performed using the fuzzy reordering metric are identical to the subjective scores for the Parsers experiment but differ on one pairwise comparison for the Rules experiment. According to BLEU score, however, none of the parser changes are significant at the 95 percent level and only one pairwise comparison (between the two most different systems) was significant at the 90 percent level. BLEU score appears more sensitive to the larger changes in the Rules experiment but is still in disagreement with the results of the human evaluation on four pairwise comparisons.

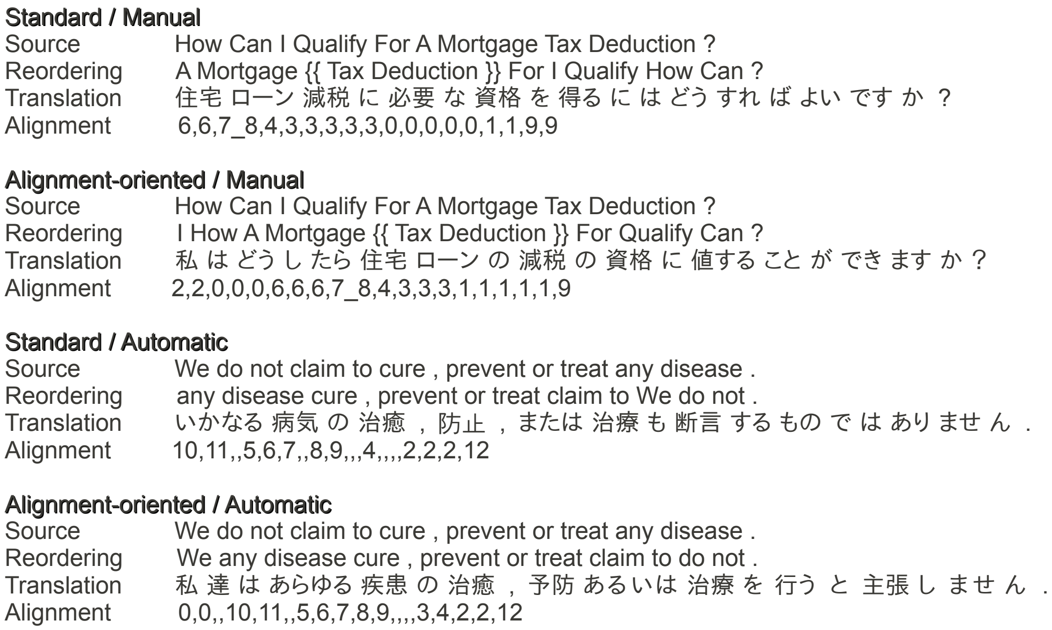
Table 6 shows the sentence-level correlation of different metrics with human judgments of translation quality. Here both the fuzzy reordering score and Kendall's τ are computed on the reference reordering data generated as described in Section 3.1. Both metrics are computed by running our lightweight evaluation tool and involve no translation whatsoever. These lightweight metrics are also more correlated with subjective quality than BLEU score at the sentence level.

Translation	Alignment	Sentence-level	
		r	ρ
Alignment-oriented	Manual	0.435	0.448
Alignment-oriented	Automatic	0.234	0.252
Standard	Manual	0.271	0.257
Standard	Automatic	0.177	0.159

Table 7: Pearson's correlation (r) and Spearman's rank correlation (ρ) with subjective translation quality at the sentence-level for different types of reordering reference data: (i) alignment-oriented translation vs. standard, (ii) manual vs. automatic alignment.

Table 7 shows how the correlation between fuzzy reordering score and subjective translation quality degrades as we move from manual to automatic alignments and from alignment-oriented translations to standard ones. The automatically aligned references, in particular, are less correlated with subjective translation scores then BLEU; we believe this may be due to the poor quality of word alignments for languages such as English and Japanese due to the long-distance reordering between them.

Finally we present some intrinsic evaluation metrics for the parsers used in the first of our experiments. Table 8 demonstrates that certain changes may not be best captured by standard parser benchmarks. While the first four parser models improve on the WSJ benchmarks as they improve subjective translation quality the best parser according to subjective translation qualtiy (P5) is actually the worst under both metrics on the treebank data. We conjecture that this is due to the fact that P5 (unlike the other parsers) is case insensitive. While this helps us significantly on our test set drawn from the web, it



Standard / Manual

Source	How Can I Qualify For A Mortgage Tax Deduction ?
Reordering	A Mortgage {{ Tax Deduction }} For I Qualify How Can ?
Translation	住宅 ローン 減税 に 必要 な 資格 を 得る に は どう すれ ば よい です か ？
Alignment	6,6,7_8,4,3,3,3,3,3,0,0,0,0,0,1,1,9,9

Alignment-oriented / Manual

Source	How Can I Qualify For A Mortgage Tax Deduction ?
Reordering	I How A Mortgage {{ Tax Deduction }} For Qualify Can ?
Translation	私 は どう し たら 住宅 ローン の 減税 の 資格 に 値する こと が でき ます か ？
Alignment	2,2,0,0,0,6,6,6,7_8,4,3,3,3,1,1,1,1,1,9

Standard / Automatic

Source	We do not claim to cure , prevent or treat any disease .
Reordering	any disease cure , prevent or treat claim to We do not .
Translation	いかなる 病気 の 治癒 ， 防止 ， または 治療 も 断言 する もの で は あり ませ ん ．
Alignment	10,11,,5,6,7,,8,9,,,4,,,,2,2,2,12

Alignment-oriented / Automatic

Source	We do not claim to cure , prevent or treat any disease .
Reordering	We any disease cure , prevent or treat claim to do not .
Translation	私 達 は あらゆる 疾患 の 治癒 ， 予防 あるいは 治療 を 行う と 主張 し ませ ん ．
Alignment	0,0,,10,11,,5,6,7,8,9,,,,3,4,2,2,12

Table 2: Reference reordering data generated via various methods: (i) alignment-oriented vs. standard translation, (ii) manual vs. automatic word alignment

	Exp. 1 Parsers					Exp. 2 Reordering Rules			
	P2R5	P3R5	P4R5	P5R5		P5R2	P5R3	P5R4	P5R5
P1R5	+**	+**	+**	+**	P5R1	+**	+**	+**	+**
P2R5		+**	+**	+**	P5R2		0	+**	+**
P3R5				+**	P5R3			+**	+**
P4R5				0	P5R4				+**

Table 3: Pairwise significance in subjective evaluation (0 = not significant, * = 90 percent, ** = 95 percent).

	Exp. 1 Parsers					Exp. 2 Reordering Rules			
	P2R5	P3R5	P4R5	P5R5		P5R2	P5R3	P5R4	P5R5
P1R5	+**	+**	+**	+**	P5R1	0	+**	+**	+**
P2R5		+**	+**	+**	P5R2		+**	+**	+**
P3R5			+**	+**	P5R3			+**	+**
P4R5				0	P5R4				+**

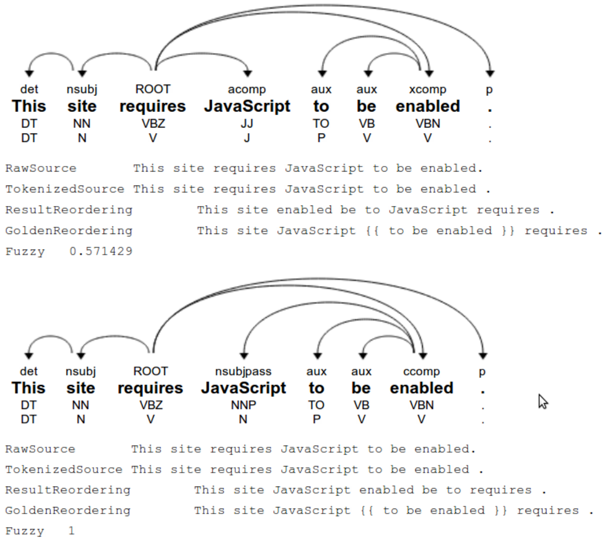
Table 4: Pairwise significance in fuzzy reordering score (0 = not significant, * = 90 percent, ** = 95 percent).

	Exp. 1 Parsers					Exp. 2 Reordering Rules			
	P2R5	P3R5	P4R5	P5R5		P5R2	P5R3	P5R4	P5R5
P1R5	0	0	+*	+*	P5R1	+**	+**	+**	+**
P2R5		0	0	0	P5R2		0	+**	+**
P3R5			0	0	P5R3			0	+*
P4R5				0	P5R4				0

Table 5: Pairwise significance in BLEU score (0 = not significant, * = 90 percent, ** = 95 percent).

Parser	Labeled attachment	POS accuracy
P1	0.807	0.954
P2	0.822	0.954
P3	0.827	0.955
P4	0.830	0.955
P5	0.822	0.944

Table 8: Intrinsic parser metrics on WSJ dev set.



```
RawSource        This site requires JavaScript to be enabled.
TokenizedSource This site requires JavaScript to be enabled .
ResultReordering        This site enabled be to JavaScript requires .
GoldenReordering        This site JavaScript {{ to be enabled }} requires .
Fuzzy   0.571429
```

```
RawSource        This site requires JavaScript to be enabled.
TokenizedSource This site requires JavaScript to be enabled .
ResultReordering        This site JavaScript enabled be to requires .
GoldenReordering        This site JavaScript {{ to be enabled }} requires .
Fuzzy   1
```

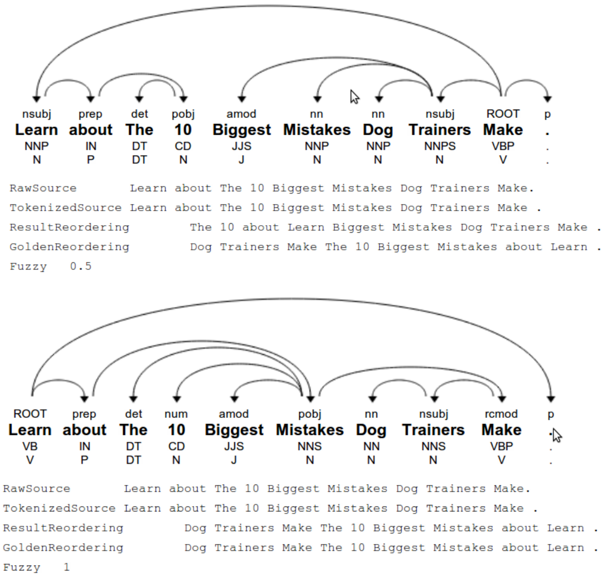
Figure 1: P1 and P5's parse trees and automatic reordering (using R5 ruleset) and fuzzy score.

hurts parsing performance on cleaner newswire.

6 Discussion

We have found that in practice this evaluation framework is sufficiently correlated with human judgments of translation quality to be rather useful for performing detailed error analysis of our English-to-Japanese system. We have used it in the following ways in simple error analysis sessions:

- To identify which words are most frequently reordered incorrectly
- To identify systematic parser and/or POS errors
- To identify the worst reordered sentences
- To evaluate individual reordering rules



```
RawSource        Learn about The 10 Biggest Mistakes Dog Trainers Make.
TokenizedSource Learn about The 10 Biggest Mistakes Dog Trainers Make .
ResultReordering        The 10 about Learn Biggest Mistakes Dog Trainers Make .
GoldenReordering        Dog Trainers Make The 10 Biggest Mistakes about Learn .
Fuzzy   0.5
```

```
RawSource        Learn about The 10 Biggest Mistakes Dog Trainers Make.
TokenizedSource Learn about The 10 Biggest Mistakes Dog Trainers Make .
ResultReordering        Dog Trainers Make The 10 Biggest Mistakes about Learn .
GoldenReordering        Dog Trainers Make The 10 Biggest Mistakes about Learn .
Fuzzy   1
```

Figure 2: P1 and P5's parse trees and automatic reordering (using R5 ruleset) and fuzzy score.

Figures 1 and 2 show pairs of parse trees together with their resulting reorderings and scores against the reference. These are typical of the parser errors that impact reordering and which are correctly identified by our framework. In related joint work (Katz-Brown et al., 2011) and (Hall et al., 2011), it is shown that the framework can be used to optimize reordering components automatically.

7 Conclusions

We have presented a lightweight framework for evaluating reordering in machine translation and demonstrated that this is able to accurately distinguish significant changes in translation quality due to changes in preprocessing components such as the parser or reordering rules used by the system. The sentence-level correlation of our metric with judgements of human translation quality was shown to be higher than other standard evaluation metrics while our evaluation has the significant practical advantage of not requiring an end-to-end machine translation experiment when used to evaluate a separate reordering component. Our analysis has also highlighted the benefits of creating focused evaluation data that attempts to factor out some of the phenomena found in real human translation. While previous work has provided meta-analysis of reordering metrics across quite independent systems, ours is we believe the first to provide a detailed comparison of systems that differ only in small but realistic aspects such as parser quality. In future work we plan to use the framework to provide a more comprehensive analysis of the reordering capabilities of a broad range of machine translation systems.

References

Alexandra Birch and Miles Osborne. 2010. Lrscore for evaluating lexical and reordering quality in mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 327–332, Uppsala, Sweden, July.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii, October. Association for Computational Linguistics.

Alexandra Birch, Phil Blunsom, and Miles Osborne. 2009. A quantitative analysis of reordering phenomena. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205, Athens, Greece, March.

Alexandra Birch, Miles Osborne, and Phil Blunsom. 2010. Metrics for mt evaluation: evaluating reordering. *Machine Translation*, 24:15–26, March.

Chris Callison-burch and Miles Osborne. 2006. Re-evaluating the role of bleu in machine translation research. In *In EACL*, pages 249–256.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005a. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005b. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 531–540, Stroudsburg, PA, USA. Association for Computational Linguistics.

Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 304–3111, July.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Comput. Linguist.*, 33:293–303, September.

Keith Hall, Ryan McDonald, and Jason Katz-Brown. 2011. Training dependency parsers by jointly optimizing multiple objective functions. In *Proc. of EMNLP 2011*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA, October. Association for Computational Linguistics.

Jason Katz-Brown, Slav Petrov, Ryan McDonald, Franz Och, David Talbot, Hiroshi Ichikawa, Masakazu Seno, and Hideto Kazawa. 2011. Training a Parser for Machine Translation Reordering. In *Proc. of EMNLP 2011*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

J. Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Chao Wang. 2007. Chinese syntactic reordering for statistical machine translation. In *In Proceedings of EMNLP*, pages 737–745.

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253, Boulder, Colorado, June.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63.

Y. Zhang and S. Clark. 2008. A Tale of Two Parsers: Investigating and Combining Graph-based and Transition-based Dependency Parsing. In *Proc. of EMNLP*.