DTED: Evaluation of Machine Translation Structure Using Dependency Parsing and Tree Edit Distance

Martin McCaffery
School of Computer Science
University of St Andrews
KY16 9SX, United Kingdom

Mark-Jan Nederhof
School of Computer Science
University of St Andrews
KY16 9SX, United Kingdom

Abstract

We present DTED, a submission to the WMT 2016 Metrics Task using structural information generated by dependency parsing and evaluated using tree edit distances. In this paper we apply this system to translations produced during WMT 2015, and compare our scores with human rankings from that year. We find moderate correlations, despite the human judgements being based on all aspects of the sentences while our metric is based only on word order.

1 Introduction

In the ever-growing field of translation metrics, a number of systems exist which attempt to provide an overall rating for a sentence. Most of these use one or more reference translations produced by a human as a gold standard. One of the earliest examples of such a metric may be BLEU (Papineni et al., 2002), using an adapted version of the well-known principle of Precision. More recently, NIST (Doddington, 2002) and Meteor (Lavie and Agarwal, 2007) have used n-gram analysis to provide similar heuristics, and many other techniques have been proposed (Dahlmeier et al., 2011; Gamon et al., 2005).

These metrics are useful when making high-level comparisons between several machine translation systems, but they offer very limited insight into the linguistic workings of the machine translation process. They can be used in automatic processes such as training systems through hill-climbing iterations, or as broad descriptions of a system's overall quality. It is however difficult to use this kind of score to gain more precise insights into a system's features; for example, different tasks may have different priorities for which errors are least desirable. Deeper analysis might also be able to pinpoint specific areas of improvement within a system. With these and other goals in mind, granular metrics have been created to evaluate individual aspects of the translated output in isolation (Zeman et al., 2011; Popović, 2011).

When developing such granular metrics, the question of which linguistic aspects of translations to focus on is far from trivial. While there has been much related discussion in the professional and educational spheres of the factors which can affect understanding of a given translation, the academic sphere has been less prolific. Nonetheless, a widely-used taxonomy on the distinct problem types which can be observed has been produced by Vilar et al. (2006), while Birch et al. (2008) investigated those which most affect overall understanding of a translation.

One of the prime factors identified by Birch et al. (2008) was word order, and metrics have been produced since then which focus on this factor (Talbot et al., 2011; Birch et al., 2010). These metrics apply various techniques, but most are based on the concept of comparing individual substrings of a source and reference sentence. While these techniques allow lightweight algorithms to produce rough scores, they ignore how the structure of a sentence can dramatically affect the impact of a mistake in ordering. For example, the mistake in the hypothesis of sentence 1 of Table 1 is much less significant than that of sentence 2, despite the latter being closer in a 'flat' judgement.

In an attempt to mitigate these problems, though without the explicit goal of focusing on word order, some work has been done using structural evaluation of sentences through dependency parsing (Gaifman, 1965). These systems either focus on applying BLEU-style n-gram matching to a tree context (Liu and Gildea, 2005; Owczarzak et al., 2007) or focus on specific relationships between

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 491–498,
Berlin, Germany, August 11-12, 2016.

	Reference	Hypothesis
1	I spoke to him there.	I spoke there to him.
2	She let it be and left.	She let it and be left.

Table 1: Example word order errors

and groupings of nodes in the trees and compare those features between hypothesis and reference trees to produce holistic judgements (Habash and Elkholy, 2008; Yu et al., 2014).

The approach of our system, named DTED (Dependency-based Tree Edit Distance), differs from existing word order literature by including dependency structures, but adds to the body of dependency-based work by focusing on node order rather than attempting to give an overall score. We work on complete dependency trees, rather than specific subsections, to produce an edit distance between the hypothesis and reference trees.

A tree edit distance is a count of the actions required to convert one ordered tree into another. In the manner of Levenshtein distances (Levenshtein, 1965) and Word Error Rate (Nießen et al., 2000), these actions are limited to Renaming, Deleting an existing node, or Inserting a new one. A number of variants on this model have been proposed, many attempting to improve the efficiency of the algorithm when applied in large-scale or high-throughput areas (Bille, 2005). The algorithm we have implemented is an extension of that proposed by Demaine et al. (2009), which is worst-case optimal, running in $O(n^3)$ time where n is the number of words in the shorter sentence.

Its output is thus a count of required modifications, which is in turn converted to a normalised score between 0 and 1. This is coupled with a weighting, indicating when aggregating scores to a system level what proportion of nodes were indicated as aligned by a preprocessing step. Our assumption is that the position of an aligned word is more reliable than an unaligned one, so when calculating corpus-wide scores we should disproportionately consider the information of those with many aligned words.

Our algorithm thus requires nothing more than the source and reference pairs, plus tools to calculate alignments and dependency trees for the chosen target language. We have used English, but the methodology would be easily applicable to any other target language for which these two components exist.

2 Related Work

2.1 Holistic metrics

Word Error Rate (Nießen et al., 2000) uses an approach closely linked to Levenshtein distances (Levenshtein, 1965), producing a straightforward count of the number of insertions, deletions and substitutions needed to convert the hypothesis into a given reference. The Position-Independent Error Rate (Tillmann et al., 1997) performs similar calculations without considering word ordering. More recently, Translation Error Rate (Snover et al., 2006) allows 'phrase shifting' of word groups together, while CDer (Leusch et al., 2006) places higher priority and level of detail on block movement calculations.

BLEU (Papineni et al., 2002) on the other hand has achieved success by directly comparing n-grams between the two sentences: it calculates a geometric mean of n-gram precisions and applies a penalty for short sentences.

A more recent and substantial metric, Meteor (Lavie and Agarwal, 2007), first applies the parameterised harmonic mean of the Precision and Recall (Rijsbergen, 1979), which measures the correctness of the individual word choices in the hypothesis sentence. It includes a second step, taking into account the ordering of those words. It does this by 'chunking' the sentences, finding the smallest number of groups of aligned words such that each contains words which are both adjacent and identical in both hypothesis and reference sentences. The ratio of the chunk count to the total number of aligned words represents the 'goodness' of the ordering, and is then multiplied with the original harmonic mean to produce a final score.

2.2 Unstructured word order systems

The standalone nature of the second phase of Meteor's pipeline means that we can use it in isolation and consider it an existing metric for word order. We have thus modified Meteor trivially to ignore the initial harmonic mean and produce only a fragmentation score; results for both this and the off-the-shelf system are reported in section 4.

Talbot et al. (2011) use a similar technique to Meteor-Frag, basing its results on the number of chunks of contiguous words aligned by a human annotator. Birch et al. (2010) provide a different approach to the problem, representing word order as mathematical permutations and counting indi-

vidual disagreements in order, and transformations required to convert one sentence into another, in a number of ways. They ignore all features of a text other than word order of aligned nodes, to produce a mathematically pure model, but sacrifice some of the less vital – but still useful – information represented by unaligned nodes and inter-word semantic relationships.

Contrary to the above metrics' focus on word order in isolation, two tools have been designed to provide a simple approximation of several error categories at once. Both Addicter (Zeman et al., 2011) and Hjerson (Popović, 2011) use comparisons of aligned words to provide a quick analysis of missing, unexpected and moved nodes.

2.3 Dependency-structured systems

While the above metrics all apply to n-grams or other unstructured representations of data, a number of proposals exist of metrics which use dependency parsing to represent sentence structure. Liu and Gildea (2005) improved on the base concept behind BLEU to calculate headword chain precision for unlabelled dependency trees, while Owczarzak et al. (2007) extend this to use labelled dependencies.

Habash and Elkholy (2008) use a different approach to dependency trees, merging n-gram precision subscores calculated similarly to BLEU with 'span-extended structural bigram precision subscores', using two methods to compare similarities between surface (flat) distances for different pairs of adjacent nodes. Yu et al. (2014) use a different approach again, considering only the reference trees' structural elements and observing, for a variety of structural segments which they consider most relevant, whether the hypothesis sentences contain the same words as those segments in the same order.

3 Metric design

3.1 Phase 1: parsing

In order to best represent the structure of the sentences we follow past examples and parse them into dependency trees. Dependency parsing has become recognised as providing a good balance between deep semantic analysis and simplicity of parsing procedure. First devised by Gaifman (1965), it uses a simplified semantic role analysis to link words by their dependency relations, providing a bare-bones structural description of the sentences, which can then be compared.

We used the dependency parsing framework provided by Python's NLTK toolkit (Bird, 2006). This in turn wraps around the Java-implemented Malt parser (Nivre, 2003).

3.2 Phase 2: tree edit

In order to produce a measure of the correctness of word order given the structural representations produced by dependency parsing, we now need to compare the structures. To do this, we use a tree edit distance algorithm, as originally put forward by Zhang and Shasha (1989). The principle behind a tree edit distance is to count the number of delete, insert and/or match (substitution) operations needed to turn one tree into another. In the version we use (Demaine et al., 2009), the 'insert' operation, whereby a node is created in one tree X to correspond to a node in tree Y, is simply represented by a 'delete' of the corresponding node in tree Y.

The most straightforward way of executing a tree edit distance is simply to give equal weighting to all operations on all nodes. This gives us a simple measure of the structural similarity of the two trees: two identical trees will have the minimum cost, namely one 'match' operation per node, while any sub-optimally placed nodes will need to be deleted and inserted elsewhere, costing 2 actions each. While other variants of DTED are available, this version - labelled 'Pure' in section 4 - has been used for both WMT2015 and WMT2016.

3.3 Phase 3: normalisation

The tree edit distance produced by the previous stage represents actions required to convert one tree into the other. We apply a simple formula to convert this count to a normalised score between 0 and 1: a more intuitive and comparable value when dealing with larger numbers of sentences. This is done slightly differently depending on the variant of DTED being used, but the score calculated by the Pure version for a given sentence pair s, with hypothesis of length n_H and reference of length n_R, is very simple. Having determined that $dist$ actions need to be performed across the trees, we say that:

$$score_s = 1 - \frac{dist}{n_H + n_R} \quad (1)$$

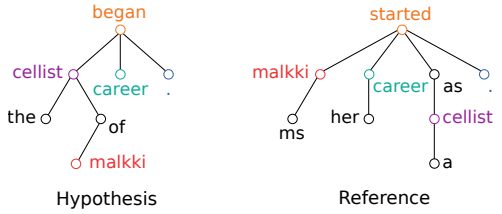

Figure 1: Sample parsed dependency trees. Matching colours show alignment between nodes.

	Reference	Hypothesis
Deleted	ms cellist a	cellist
Matched	began the of malkki . career	started malkki career her . as

Table 2: Edit operations calculated by DTED for sentences shown in Figure 1.

3.4 Variants

While we have implemented a number of variants of DTED, for WMT 2015 and WMT 2016 we use only the 'Pure' version which processes only the structure of the sentences while ignoring other information such as word alignments. DTED has been run on two types of input. First, we have run each version on normal dependency trees, leveraging the full structural information available. For comparison, we have also run each on flattened trees from which the structural information has been removed. This is done in a preprocessing step by artificially forcing each node to be the only child of its predecessor. This version is intended to nullify the structural advantage given by the rest of the system, to provide a baseline for comparison.

With the 'pure' version of DTED, the modifications shown in Table 2 are calculated.

3.5 Result aggregation

Combining individual sentence scores to an overall system-level result is done in two ways. The straightforward way is to simply take an arithmetic mean of all sentence scores, indicated in table 4 as unweighted or *not W*. This gives a total score for corpus c containing N sentences as:

$$unweighted_c = \frac{\sum_s score_s}{N} \quad (2)$$

Additionally, to investigate the importance of the aligned words in our sentence, we produce a version which assigns each sentence a weight equal to the proportion of nodes aligned in the sentence. With n_a aligned nodes and n_{na} unaligned nodes, the weight for sentence s for the Weighted version of DTED (*W* in table 4) is calculated as:

$$weight_s = \frac{n_a}{n_a + n_{na}} \quad (3)$$

For an individual sentence the score and weight can be viewed separately, while overall values for a corpus are calculated as:

$$weighted_c = \frac{\sum_s (score_s \times weight_s)}{\sum_s weight_s} \quad (4)$$

3.6 Example

Figure 1 shows dependency trees for the following sentences which occur in the WMT 2015 corpus. All pairs of words shared by both sentences are aligned, as are 'started' and 'began'.
Hyp: The cellist of Malkki began career.
Ref: Ms Malkki started her career as a cellist.

This comes to a total of 4 Delete operations and 6 Match operations, resulting in a total matching $dist$ of 10. With the hypothesis tree containing $n_H = 7$ nodes and the reference containing $n_R = 9$, we can normalise this (as per equation 1) to:

$$score_s = 1 - \frac{10}{7+9} = 0.375$$

Finally, we may optionally consider a weighting for the sentence as per equation 3.

$$weight_s = \frac{10}{10+6} = 0.625$$

This weighting indicates that we consider our low rating of the sentence partially trustworthy relative to others in the corpus.

4 Results & Discussion

4.1 Setup & Evaluation

DTED has been run on sentences provided for the 2015 (Bojar et al., 2015) and 2016 Workshops on Statistical Machine Translation. The results

Metric	cs-en	de-en	fi-en	fr-en	ru-en	all
BLEU	**0.989**	0.836	0.920	0.970	0.643	0.622
WER	0.913	0.813	0.794	0.972	0.700	0.524
TER	0.929	0.822	0.846	0.975	0.712	0.563
PER	0.980	0.764	0.858	0.967	0.753	**0.670**
CDER	0.955	0.813	0.944	**0.981**	0.762	0.561
Meteor	0.984	**0.934**	**0.961**	0.968	**0.877**	0.647

Table 3: System-level correlations of holistic metrics with normalised human rankings

Metric	*Version*	*W*	*F*	cs-en	de-en	fi-en	fr-en	ru-en	all
Meteor	Frag	-	-	0.905	0.853	**0.941**	0.927	0.781	**0.615**
DTED	Pure	X	X	0.974	**0.877**	0.841	**0.993**	**0.824**	0.522
DTED	Pure	X	✓	0.964	0.542	0.867	0.729	0.431	0.461
DTED	Pure	✓	X	**0.975**	0.872	0.814	0.992	0.822	0.522
DTED	Pure	✓	✓	0.963	0.507	0.886	0.476	0.337	0.445

Table 4: System-level correlations of word order metrics with normalised human rankings

for 2015 data are provided in this paper, while for 2016 the reader is referred to the Findings of the 2016 Workshop on Machine Translation. For the latter, DTED uses unflattened trees, without weighting by aligned nodes. Sentences from all available into-English corpora were used, but only segments for which corresponding human judgements were available. The number of individual systems for each language pair, and the count of sentences within each, are given in table 5.

Human judgements during the Workshop were given as rankings between up to 5 systems, with ties allowed. We have normalised these ranks into scores out of 1: for example, a rank of 3 between five systems is converted to 0.5, reflecting that an equal number of systems were preferred to it as were considered less good, while a system ranked best would achieve a perfect score of 1.

It should be noted that while DTED is intended to evaluate word order in isolation, rankings at WMT were based on *all* features of the sentences. As no data of sufficient quantity and quality was available for human judgements specifically of word order, we have used the holistic data. As such, we do not expect cutting-edge correlational values for this data; instead, such comparisons are provided for two separate reasons.

First, as word order is clearly involved in some non-trivial way in human judgements, we can assume that holistic ranks contain an implicit word order component. A limited level of similarity between human judgements and DTED is thus to be expected, as they are at least partially measuring the same phenomenon. In addition, while the DTED algorithm is intended to measure word order alone, the structure and alignment of the trees we use may themselves depend on other factors. For example, a badly chosen word may occupy a different role in its sentence than the reference choice would, resulting in an unpredictable change in the actions needed to correct it.

Second, if we assume DTED's results to be successfully representative of a sentence's word order quality and human judgements to contain a word order component, the level of correlation can begin to quantify the significance of word order within the overall judgement. In the ideal theoretical case where DTED perfectly simulated human intuition on word order, such correlational coefficients would give direct insight into the significance of that intuition to overall quality judgements.

4.2 Ratings

We have performed analysis on two types of metric: holistic and word order specific. Table 3 compares human judgements to those produced by a number of well-known and widely used baseline metrics, while table 4 shows the same values for metrics designed specifically for word order. In both tables, the highest score for each corpus is highlighted.

Meteor's fragmentation-only subsystem (see section 2.2) is included in the latter table, while

	cs-en	de-en	fi-en	fr-en	ru-en	all
Num. systems	7	13	14	6	13	53
Total sentences	909	692	510	815	782	3708

Table 5: Sizes of corpora used for all empirical calculations, all produced during WMT 2015

the version of Meteor in the former is a standard off-the-shelf installation. For DTED, the *W* column indicates whether sentences were considered equally when aggregating, or were *W*eighted based on aligned word content as per section 3.5. Results run on *F*lattened trees (section 3.4) are indicated by the column *F*.

All scores except those for DTED and Meteor were calculated using implementations of the metrics provided with the well-known open-source system Moses (Koehn et al., 2007). In all cases, the numbers shown are Pearson correlation coefficients between the output of the given metric at the system level and the normalised human judgements provided at WMT 2015.

4.3 Discussion

The main trend we can see from tables 3 and 4 is that for the versions of DTED with the highest correlation values to human judgement, those values are similar to, if marginally lower than, the scores of the baseline metrics. To represent this trend, the unflattened version of DTED (irrespective of weighting) has an overall correlation almost exactly the same as the baseline metric WER which performed the most poorly.

While the correlations of DTED versions are thus fairly encouraging when compared to those of other metrics, they are also interesting when compared to each other. An almost universal trend is that when applied on flattened trees DTED was significantly less effective in predicting human judgements. This strongly indicates that we have succeeded in leveraging the structural information in the non-flattened dependency trees and used the information to good purpose in a similar way to a human.

It should be noted that weighting the sentences according to the proportion of aligned nodes provided a boost to correlations, albeit an extremely small one.

5 Conclusions & Future Work

DTED represents the first work we know of which uses tree edit distances to incorporate structure into the evaluation of machine translation word order. Our results suggest that this approach, while not as holistically accurate as metrics designed for that purpose, nonetheless provides scores with non-trivial similarities to human ratings. This suggests that our metric does indeed measure a significant component of humans' intuition on sentence quality for English. While not a conclusion that can be drawn from the empirical results as such, we feel confident that our metric does primarily evaluate word order as opposed to other factors such as word choice. Taking these two assumptions together, we can say that a significant component of humans' sentence-quality intuition is based on the order of words.

Though the statement that word order accounts for a large part of humans' quality judgements is highly interesting, it would be worthwhile to investigate the relationship more directly. An obvious way to produce results more tailored to it would be to obtain human judgements based solely and explicitly on word order. Such judgements would also allow us to more appropriately evaluate the more alignment-focused versions of DTED: while in the experiments we have performed on WMT judgements these have done less well, this may simply be because these variants are intended to more precisely focus on word order. An increase in such precision will necessarily result in less broad scores and thus lower correlation with the broad-scope judgements available.

While tree edit distance leverages much of the information contained in structural representations of sentences, it fails to account for the distances through which nodes must be moved. We thus intend to consider models more akin to gradual movement than disparate operations, such as those related to the concept of inversion numbers (Conlon et al., 1999). A further avenue of investigation would be whether the structural and order-specific functionality of a tree edit distance could be approximated or reproduced by a more lightweight algorithm.

References

Philip Bille. 2005. A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337(1-3):217–239, jun.

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

Alexandra Birch, Miles Osborne, and Phil Blunsom. 2010. Metrics for MT evaluation: evaluating reordering. *Machine Translation*, 24(1):15–26, jan.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions (COLING-ACL '06)*, pages 69–72.

Ondej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, pages 1–46, Lisboa, Portugal.

Margaret M Conlon, Maria Falidas, Mary Jane Forde, John W Kennedy, S McIlwaine, and Joseph Stern. 1999. Inversion numbers of graphs. *Graph Theory Notes of New York*, 37:42–48.

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. 2011. TESLA at WMT 2011: Translation evaluation and tunable metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 78–84.

Erik D. Demaine, Shay Mozes, Benjamin Rossman, and Oren Weimann. 2009. An optimal decomposition algorithm for tree edit distance. *ACM Transactions on Algorithms*, 6:1–19.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145.

Haim Gaifman. 1965. Dependency systems and phrase-structure systems. *Information and Control*, 8(3):304–337.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proceedings of the European Association for Machine Translation*, pages 103–111.

Nizar Habash and Ahmed Elkholy. 2008. SEPIA: Surface Span Extension to Syntactic Dependency Precision-based MT Evaluation. In *Proceedings of the NIST Metrics for Machine Translation Workshop at the Association for Machine Translation in the Americas Conference*, Waikiki, HI.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej 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*, pages 177–180. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDer: Efficient MT evaluation using block movements. In *Proceedings of EACL-2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, pages 241–248.

Vladimir Iosifovich Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Dokl.*, 10(1):707–710.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Sonja Nießen, Franz-Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: fast evaluation for MT research. *LREC*, pages 0–6.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. *Proceedings of the 8th International Workshop on Parsing Technologies*, pages 149–160.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Labelled dependencies in machine translation evaluation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 104–111.

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.

Maja Popović. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–67.

C J Van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Matthew Snover, Bonnie Dorr, College Park, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,*, number August, pages 223–231, Cambridge, Massachusetts.

David Talbot, Hideto Kazawa, Hiroshi Ichikawa, Jason Katz-Brown, Masakazu Seno, and Franz-Josef Och. 2011. A lightweight evaluation framework for machine translation reordering. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 12–21. Association for Computational Linguistics.

C Tillmann, S Vogel, H Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP Based Search for Statistical Translation. *Fifth European Conference on Speech Communication and Technology*, pages 2667–2670.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 697–702, Genoa.

Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2014. RED: A Reference Dependency Based MT Evaluation Metric. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2042–2051.

Daniel Zeman, Mark Fishel, Jan Berka, and Ondej Bojar. 2011. Addicter: What is Wrong with My Translations? *The Prague Bulletin of Mathematical Linguistics*, (96):79–88.

Kaizhong Zhang and Dennis Shasha. 1989. Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM Journal on Computing*, 18(6):1245–1262.