Metrics for MT Evaluation: Evaluating Reordering

Alexandra Birch (a.c.birch-mayne@sms.ed.ac.uk) and
Miles Osborne (miles@inf.ed.ac.uk)
University of Edinburgh

Phil Blunsom (phil.blunsom@comlab.ox.ac.uk)
University of Oxford

Abstract. Translating between dissimilar languages requires an account of the use of divergent word orders when expressing the same semantic content. Reordering poses a serious problem for statistical machine translation systems and has generated a considerable body of research aimed at meeting its challenges. Direct evaluation of reordering requires automatic metrics that explicitly measure the quality of word order choices in translations. Current metrics, such as BLEU, only evaluate reordering indirectly. We analyse the ability of current metrics to capture reordering performance. We then introduce permutation distance metrics as a direct method for measuring word order similarity between translations and reference sentences. By correlating all metrics with a novel method for eliciting human judgements of reordering quality, we show that current metrics are largely influenced by lexical choice, and that they are not able to distinguish between different reordering scenarios. Also, we show that permutation distance metrics correlate very well with human judgements, and are impervious to lexical differences.

Keywords: Machine Translation, Metrics, Reordering, BLEU, METEOR, TER, Permutation Distances, Human Evaluation

1. Introduction

Relative grammatical properties of human languages can result in a wide variety of word order differences when performing translation. The word order differences are difficult to model and the amount of reordering has been shown to be a very important predictive factor in translation performance (Birch et al., 2008). We argue that research in reordering is hampered by the fact that commonly used automatic metrics do not explicate the reordering performance of machine translation systems. A reliable reordering metric is a key resource for further improvements in the field.

Current automatic metrics only measure word order differences indirectly. They are generally sensitive to lexical differences but this affects their ability to determine the correct word order. None of the metrics take the size of the word order differences into account. Furthermore, they all have parameters which are difficult to train and further obscure the reordering component of the rating. Although reordering improvements can be linked with improvements in lexical choice, evaluating reordering largely on the basis of differences in lexical choice is not

satisfactory. We evaluate current automatic metrics to see how well they measure word order differences.

We argue that it is important to evaluate reordering performance directly. We present a method for doing this, in isolation from lexical choice, which uses **permutation distance metrics**. The assumption is that orderings that are close to that of the reference word order are going to be preferable to orderings which are very different. We first extract permutations from alignments, and then apply standard distance metrics. Distance metrics are intuitive measurements that are sensitive to the size and frequency of reorderings. They are also efficient, language independent and they are meaningful at a sentence level. These properties make them desirable automatic machine translation metrics.

Automatic metrics need to be validated by human judgements. We develop a novel human evaluation task which specifically measures reordering performance. Correlation between the automatic metrics and human judgement shows that all metrics measure reordering success where there is a perfect lexical match between the reference and the test sentence. The Hamming distance, Kendall's tau distance and the METEOR score correlate the best. However, for experiments on real translations where the lexical overlap is reduced, the MT metrics BLEU, METEOR and TER are shown to be remarkably insensitive to reordering differences, and they are mainly influenced by the quality of word choice.

2. Evaluating Metrics

Good automatic metrics of translation quality are key to developing better machine translation systems. There is a lot of interest in both developing metrics and in their evaluation, as is shown by recent evaluation campaigns. The Workshop on Statistical Machine Translation (Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009) and the NIST Metrics for Machine Translation 2008 Evaluation[1] have both collected human judgement data to evaluate a wide spectrum of metrics. However, there has been very little research which has specifically addressed reordering.

We analyse three current MT metrics, the BLEU score (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and the Translation Edit Rate (TER) (Snover et al., 2006). These three metrics have all performed well in the evaluation campaigns and they are widely used. They are all shallow metrics - no deep linguistic analysis is required. This is important as it makes them more language independent and

[1] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/

faster to compute making them more widely usefull and appropriate for training systems. Other metrics which demonstrate good correlation with human ratings, such as ULC (Giménez and Màrquez, 2007) and Rich Textual Entailment (Padó et al., 2009) both combine simple and more complex features such as semantic and dependency overlap.

BLEU is the de facto standard in machine translation. It captures the n-gram precision of the translation. Shorter n-gram precision captures the lexical coverage of the translation and word order is evaluated by the higher order n-grams. The final score is an interpolation of these precisions and it is adjusted by a brevity penalty. There is no intuitive meaning behind a BLEU score, and BLEU scores on different language pairs or corpora cannot be compared. Another disadvantage is that the same BLEU score has no measure of how far words are reordered (Callison-Burch et al., 2006). This is a common problem with the other automatic metrics. METEOR measures precision and recall for unigrams, it is able to match words with common stems or synonyms. However, its strength at matching lexical items is balanced by using a poor measure of word order similarity. It applies a fragmentation penalty which counts the number of chunks the translation needs to be broken into, to align to the reference. The TER score measures the number of edits required to change the hypothesis into the reference. The edit operations are insertion and deletion of words, and moves of contiguous chunks. The 'move' operation relies upon an exact lexical match, and does not consider the size of the reordering. All these metrics have parameters which are difficult to train, and make the interpretation of the score more difficult. None of these metrics have been directly evaluated on a reordering task.

3. Permutation Distance Metrics

The reorderings of a sentence pair can be encoded as a **permutation**, allowing us to apply research into distance functions for ordered encodings to our task of evaluating reorderings. A word alignment over a sentence pair allows us to transcribe the source word positions in the order of the aligned target words. The permutation from the source-reference alignment is then compared with the permutation of the source-translation alignment. Where accuracy is paramount, gold standard human annotated alignments can be used.

Permutation distance metrics present many of the qualities of the ideal metrics. They are intuitive because they are distance metrics, having nice properties such as equality, symmetry and triangle inequality. The metrics are language independent as they only depend on alignments, not on lexical choice. They are sentence level metrics which are fast to calculate. These metrics explicitly measure reordering, some-

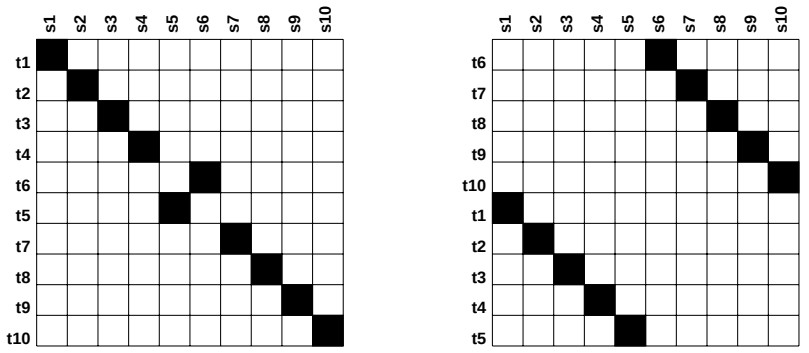


Figure 1. Synthetic examples of two example alignments and their permutations where: (a) there is one short distance word order difference, (b) the order of the two halves has been swapped. Bullet points represent the non-sequential gaps in the permutation.

thing which other translation metrics do not do. They are extensible to multiple references, either by averaging the metrics for each reference, or by selecting the most appropriate reference to compare with. Permutation distance metrics have been used to evaluate data in natural language processing tasks before. Kendall's tau (Lapata, 2003) was used as a means of estimating the distance between a system-generated and a human-generated gold-standard order for the sentence discourse task. The correlation between Kendall's tau and human judgements was also established (Lapata, 2006).

A permutation is a **bijective function** from a set of natural numbers $1, 2, \cdots, N$ to itself. We will name our permutations π and σ. The ith symbol of a permutation π will be denoted as $\pi(i)$ and the identity, or monotone, permutation id is the permutation for which $id(i) = i$ for all i.

Figure 1 shows an example of two sentence pairs and their permutations. The permutations are calculated by iterating over the source words (s1...s10), and outputting the position of the aligned target words (t1...t10). Alignments can represent situations that are more complex than permutations can account for and we make some simplifying assumptions. Phrasal alignments are assumed to be monotone word alignments. Non-contiguous alignments are simplified to only record the first target word position. Null source word alignments are assumed to align to the previous word.

Permutation distance metrics are calculated by comparing permutations extracted from a source sentence aligned to different target sentences. We invert the distance metrics by subtracting from one, so

that an increase in the metrics represents an increase in the quality of word order. We define and discuss the different metrics below.

– The **Hamming Distance** measures the number of disagreements between two permutations (Ronald, 1998):

$$d_H(\pi,\sigma) = 1 - \frac{\sum_{i=1}^{n} x_i}{n} \text{ where } x_i = \begin{cases} 0 \text{ if } \pi(i) = \sigma(i) \\ 1 \text{ otherwise} \end{cases} \quad (1)$$

Where π, σ are the two permutations and n is the length of the permutation. The Hamming distance captures the amount of absolute disorder that exists between two permutations.

– The **Ulam's Distance** is the minimum number of single item movements of arbitrary length required to transform one permutation into another (Ulam, 1972). It is defined as follows:

$$D_U(\pi,\sigma) = 1 - \frac{n - LCS(\pi,\sigma)}{n} \quad (2)$$

where LCS is the longest common subsequence. A subsequence is a sequence that can be derived from another sequence by deleting some entries without changing the order of the remaining entries. For example, (2 3 5 7) is a subsequence of (1 2 3 4 5 6 7). The LCS can be calculated in $O(n^2)$ by a simple dynamic programming algorithm (Hirschberg, 1975). The Ulam distance is not a measure of absolute order. It is sensitive to the relative order of words, but only to those that fall within the LCS.

– The **Kendall's Tau Distance** is the minimum number of transpositions of two *adjacent* symbols necessary to transform one permutation into another (Kendall and Dickinson Gibbons, 1990):

$$d_\tau(\pi,\sigma) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} z_{ij}}{Z} \quad (3)$$

$$\text{where } z_{ij} = \begin{cases} 1 \text{ if } \pi(i) < \pi(j) \text{ and } \sigma(i) > \sigma(j) \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

$$Z = \frac{(n^2 - n)}{2} \quad (5)$$

It reflects the sum of all pairwise differences in order between the two permutations. The Kendall's tau metric is sensitive to relative orderings between all pairs of words and therefore to the distance

that words are reordered. It can be interpreted as the probability that pairs of items in two different permutations are in the same order as opposed to being in different orders.

Table I. Metric scores for examples in Figure 1 which are calculated by comparing the permutations to the identity.

Example	BLEU	METEOR	TER	D_H	D_U	$D_{K\tau}$
(a)	61.80	86.91	90.00	80.00	90.00	97.77
(b)	81.33	92.63	90.00	0.00	50.00	44.44

In order to highlight the problem of the current MT metrics, and to demonstrate how the permutation distance metrics are calculated, in Table I we present the metric results for the two example sentence pairs from Figure 1. The MT metrics are calculated by comparing their target string with the monotone target string (t1...t10). For now, we assume that the reference and the translation contain exactly the same words, only they are ordered differently. The permutation distance metrics only consider the ordering of the target string and compare it with the identity permutation (1...10). In order to facilitate comparison, we subtract all distance metrics (including TER) from one, and all scores are reported as percentages.

The example sentence pair (a) represents a small reordering and (b) a large one, however, most of the MT metrics fail to recognise this and they assign a poorer score to (a). The reason for this is that they are sensitive to breaks in order, but not to the actual word order differences. The BLEU score detects three breaks in order in example (a) with a score of 61.80, but only one break for (b). It therefore matches more n-grams for (b) and consequently assigns a higher score of 81.33. METEOR counts the number of blocks that the translation is broken into, in order to align it with the source. (a) is aligned using four blocks and scores 86.91, whereas (b) is aligned using only two blocks and scores 92.63. TER counts the number of edits, allowing for block shifts. TER applies one block shift for each example, resulting in and equal score for both sentences of 90.00 and thus demonstrating its insensitivity to the amount of reordering.

The reordering metrics correctly assign a lower score to (b) as they recognise the number of words affected by reordering. The Hamming distance detects two words out of order in (a), resulting in a score of $(1 - \frac{2}{10}) * 100 = 80$, and all words out of order in (b) giving the worst score of 0. The Ulam distance detects a LCS of 9 for the example (a) giving a score of $(\frac{9}{10}) * 100 = 90$, but for (b) there is only an LCS of 5, and thus a score of 50. Kendall's tau metric also takes the distance

words have moved into account. For example (a) the metric detects only one pair of words out of order $(1 - \frac{1}{45}) * 100 = 97.77$ whereas for example (b) there are 25 and a score of $(1 - \frac{25}{45}) * 100 = 44.44$.

4. Correlation with human judgements of reordering

To assess whether automatic metrics reliably correlate with human ratings, we present an experiment which evaluates several different orderings of the same input. In what follows we describe our method for assembling a set of experimental materials and collecting human judgements.

4.1. Design and Materials

The question of how best to use humans to evaluate machine translation output is still unresolved. Most human evaluations are performed on the output of translation systems, and therefore are subject to many confounding factors such as sentence difficulty, lexical choice, domain etc. In order to isolate word order differences as the only factor affecting human judgements, we construct a controlled experiment. This experiment is based on a set of sentences which have been translated and word aligned by human annotators. The English reference sentence word order is then scrambled to different degrees creating versions with different amounts of reordering as compared to the original. Users are presented with one version of each sentence and asked to rate them for fluency and comprehension.

We use the Chinese-English parallel corpus that are provided by the GALE project[2] as it provides human annotated word alignments. We randomly select 40 sentences which have a large amount of reordering (RQuantity (Birch et al., 2008) > 1.3) and where the sentence length is between 10 and 40 words.

There are five versions of each test sentence and each falls into a bin with a different amount of reordering or RQuantity. Bin 0 contains the reference. Bin 4 contains the reference in the Chinese word order. Bins $1 \ldots 3$ contains three intermediate versions. The English word order of each intermediate version is the result of applying a random subset of the reorderings that were detected in the original Chinese-English sentence pair. We choose to explore this particular space of possible word orders because it represents a wide range of humanly plausible reorderings. If we had explored the space of orderings that a translation system could produce, this would only represent a small and biased range of orderings. If, on the other hand, we had chosen to

[2] see LDC corpus LDC2006E93 version GALE-Y1Q4

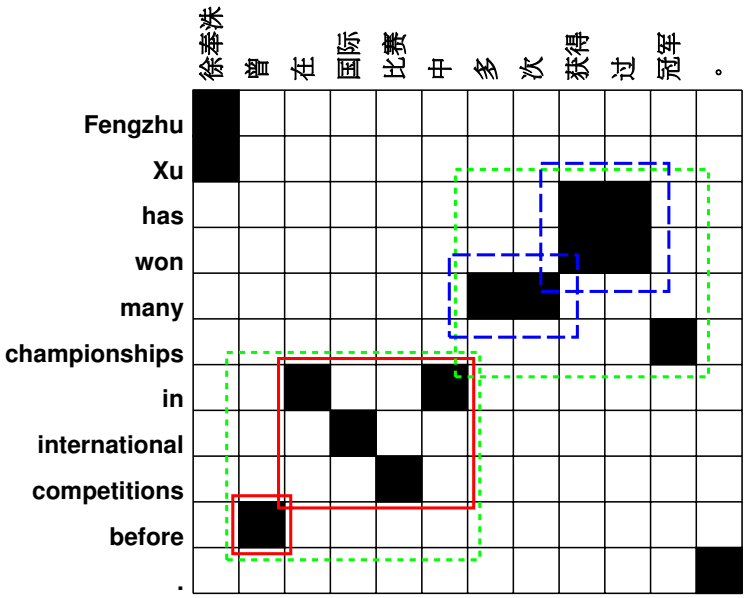
represent all possible word orderings, from inverted to monotone, this would represent a vast and totally implausible set of reorderings.

Figure 2 shows an example from the experiment. The original sentence pair in (a) is shown together with the shuffled test cases in (b) and their scores on the metrics in (c), and the resulting averaged human evaluations. Test sentences in Bins 1, 2, and 3 were created by applying a random subset of the reorderings shown in (a). For the case of the test sentence in Bin 1, the movement of the word 'before' is due to retaining all but the small reordering shown in the solid line rectangles (with blocks 'in international competitions' and 'before'). Likewise with the test sentence in Bin 3, where the disorder is due to retaining all but the large reordering shown in the short-dash rectangles (with blocks 'has won many championships' and 'in international competitions before'). The final test case in Bin 4 is the English sentence as read out using the Chinese word order.

This example again highlights problems with the MT metrics and show that humans agree with reordering metrics, at least in fluency judgements. Bin 2 has two small reorderings and when compared to Bin 3 it gets on average a one point higher fluency rating from judges. Reordering metrics also give a higher score to Bin 2, but MT metrics give a higher score to Bin 3 because they are not sensitive to the size of the reordered chunks. Looking at Bin 4, we can see that it is completely garbled, but interestingly, the average humans judgement of comprehension is quite high. That might be because it is a short sentence and for simple sentences the meaning can be guessed.

The study was conducted remotely over the Internet using Webexp[3]. Subjects were shown instructions which included examples of well and badly ordered sentences. From our set of sentences we created five lists each consisting of 40 sentences, following a Latin square design. Each subject was randomly assigned one list which ensured that no user saw more than one version of the same sentence. For each test case the subject was asked for a judgement on its fluency and adequacy on a seven point scale. The experiment was completed by 28 unpaid volunteers, all self-reported fluent speakers of English. We used the smoothed BLEU metric (Lin and Och, 2004) to calculate BLEU on a sentence level. These measurements will be less stable than BLEU on a document level, but they are a reasonable approximation. We selected the stemming and synonymy modules to use with METEOR, even though all the words in the two sentences are identical and they should not be necessary.

[3] http://www.webexp.info/



(a)

Bin	Test cases with Permutations
Ref	Fengzhu Xu has won many championships in international competitions before . 1 2 3 4 5 6 7 8 9 10 11
1	Fengzhu Xu has won many championships before in international competitions . 1 2 3 4 5 6 • 10 • 7 8 9 • 11
2	Fengzhu Xu many has won championships before in international competitions . 1 2 • 5 • 3 4 • 6 • 10 • 7 8 9 • 11
3	Fengzhu Xu in international competitions before has won many championships . 1 2 • 7 8 9 10 • 3 4 5 6 • 11
4	Fengzhu Xu before in international competitions many has won championships . 1 2 • 10 • 7 8 9 • 5 • 3 4 • 6 • 11

(b)

Bin	BLEU	MET.	TER	D_H	D_U	$D_{K\tau}$	Fluency	Compr.
Ref	100.00	100.00	100.00	100.00	100.00	100.00	6.43	6.71
1	66.36	89.69	90.90	63.63	90.90	94.54	6.00	6.57
2	31.70	81.67	81.81	36.36	81.81	90.90	5.25	6.50
3	59.00	89.69	90.90	27.27	63.63	70.90	4.25	6.50
4	31.70	81.67	72.72	27.27	54.54	61.81	2.28	5.57

(c)

Figure 2. An example of a sentence pair used in the human evaluation campaign. The sentence pair in (a) is shown with the alignment and the reorderings, displayed with rectangles of different line styles. In (b) are displayed the five differently ordered test cases that are presented for evaluation to the human judges. Finally at the bottom in (c), a table with scores for the different test versions are presented, including metric scores and resulting average human judgements on fluency and comprehension.

4.2. Results

We use correlation analysis to explore the linear relationship between human judgements the permutation distance metrics. This shows us if they are indeed appropriate measures for evaluating reordering, and which metrics are best at capturing the reordering differences. In Table II we see the Pearson's correlation coefficients for the metrics compared to the human fluency and comprehension ratings, averaged per test item. All the correlations are significant to the 99.9% level.

We can see that all correlations are quite strong, with the strength of the correlation for comprehension generally being lower than that of fluency. This is explained by the fact that sometimes a sentence can be disfluent but one can still make out its meaning. Kendall's tau correlates slightly less than the other metrics. We analysed the data for Kendall's tau distance and discovered that our assumption of a linear relationship between human judgements and Kendall's tau was incorrect. A linear model fits much better after transforming the Kendall's tau values by taking the square root. This new measure also correlates very well with human judgements. It seems that because Kendall's tau is sensitive to the distance that a word is reordered, it is mainly influenced by very large reorderings that are not commonly seen in human translations. The reorderings seen in our humanly plausible reordering experiment result in quite small Kendall's tau values, see Figure 2. It seems that more sensitivity to the smaller reorderings is needed, and this is why taking the square root seems to fit human judgements better.

The strength of the Hamming distance is somewhat surprising as it is a very simple metric that depends on absolute position. If the first word is reordered to the end of a sentence, the Hamming distance will be zero because the fact that all the other words retain their relative order is not taken into account. However, the Hamming distance seems to be more informative than Kendall's tau for small amounts of reordering.

5. Factors influencing metrics

In previous experiments we used human judgements derived from an artificial experiment to evaluate the metrics. We have seen that under artificial test conditions where there is perfect lexical overlap between the reference and the translation, current MTmetrics correlate reasonably well with human judgements. However, MTmetrics are expected to perform much worse where there is lexical variation between reference and translation. We designed an experiment to analyse what contribution lexical variation and word order performance have on the variability of the current MTmetrics, under real test conditions.

Table II. The Pearsons correlation of metrics with averaged human fluency and comprehension judgements.

Metric	*Fluency*	*Comprehension*
BLEU	0.779	0.619
METEOR	0.802	0.638
TER	0.712	0.602
Hamming	0.806	0.664
Ulam	0.766	0.656
Kendall's τ	0.707	0.599
$\surd$ Kendall's τ	0.795	0.656

While the permutation distances are insensitive to lexical differences, the ability of MT metrics to detect word order differences is hampered by differences in word choice. BLEU will consider every non-matching word to be a break, and so ordering differences will only be detected if they occur between words which are identical in the translation and the reference. METEOR will try to match synonyms and stems which leads to errors in the alignment. TER can account for differences in word order by using insertions and deletions. All MT metrics conflate the lexical and the ordering component of the measure, making it difficult to know what the actual reordering performance is.

We used the 1-gram BLEUscore as our measure of lexical overlap. This is a precision score which takes into account multiple reference sentences and is defined as the number of matched words divided by the length of the translation. We have demonstrated that we are able to capture the reordering performance of sentences using the adjusted Kendall's tau distance, which measures relative order and takes the size of reorderings into account. Multiple references are accounted for by measuring the distance to the reference with the closest word order.

The test data comprised of 1994 sentences from the GALE 2008[4] Chinese-English newswire test set which each have four English reference sentences. We trained a phrase-based model using MOSES (Koehn et al., 2007) on the full GALE 2008 Chinese-English training corpus. With all the default options, we generated the translation output in English from the Chinese source sentence. We then word aligned the reference and the translated sentences to the Chinese source using the Berkeley word aligner (Liang et al., 2005) which was also trained on the full GALE 2008 training corpus.

[4] http://www.itl.nist.gov/iad/mig/tests/gale/2008/

Table III. The Pearson's correlations and the R^2 of simple linear regression models exploring the size of the correlation between lexical choice and reordering on the current metrics. All regressions are significant to the 99.9% level

Pearsons Correlation			Linear Regression R^2		
Metric	*BLEU1*	$\sqrt{}$ *Kendall's* τ	*Metric*	*BLEU1*	$\sqrt{}$ *Kendall's* τ
BLEU	0.693	0.255	BLEU	0.481	0.065
METEOR	0.609	0.162	METEOR	0.371	0.026
TER	0.736	0.302	TER	0.543	0.091

We perform correlation and regression analysis on MT metrics and present the results in Table III. The Pearson's correlation results show that there is a much stronger correlation between lexical overlap and the MT metrics than between reordering and the MT metrics. Although correlation is a good indication of the strength of the relationship, regression and the R^2 statistic allow us to describe the proportion of variance in the dependent variable that can be accounted for by the regression equation. Here we can see that reordering seems to have a minimal effect on all of the metrics. This leads us to conclude that current metrics BLEU, METEOR and TER are on the whole largely insensitive to reordering differences and are mainly affected by lexical choice. Finally, it is interesting to note that TER is more correlated with both lexical choice and reordering than the other two metrics, and METEOR is less correlated than the other two metrics.

6. Conclusion

We evaluated current metrics for their ability to measure reordering performance. We also proposed measuring reordering explicitly by using permutation distance metrics, which have some nice properties. A human experiment was devised where reordering was isolated for extracting fluency and comprehension judgements. Comparing with human reordering judgements, we found that all metrics correlated strongly with fluency judgements, but that the square root of Kendall's tau distance was the best metric, because it was more reliable than Hamming distance and correlated almost as strongly. Current metrics were found to be largely influenced by lexical choice and insensitive to reordering differences. In the future we plan to develop a combined lexical and reordering metric which could be used instead of current MT metrics.

References

Birch, A., M. Osborne, and P. Koehn: 2008, 'Predicting Success in Machine Translation'. In: *Proceedings of the Empirical Methods in Natural Language Processing*.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder: 2007, '(Meta-) Evaluation of Machine Translation'. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 136–158.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder: 2008, 'Further Meta-Evaluation of Machine Translation'. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, pp. 70–106.

Callison-Burch, C., P. Koehn, C. Monz, and J. Schroeder: 2009, 'Findings of the 2009 Workshop on Statistical Machine Translation'. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, pp. 1–28.

Callison-Burch, C., M. Osborne, and P. Koehn: 2006, 'Re-evaluation the Role of BLEU in Machine Translation Research'. In: *Proceedings of EMNLP*.

Diaconis, P. and R. L. Graham: 1977, 'Spearman's Footrule as a Measure of Disarray'. *Royal Statistical Society Series B* **32**(24), 262–268.

Giménez, J. and L. Màrquez: 2007, 'Linguistic Features for Automatic Evaluation of Heterogenous MT Systems'. In: *ACL Workshop on Statistical Machine Translation*.

Kendall, M. and J. Dickinson Gibbons: 1990, 'Rank Correlation Methods'. New York: *Oxford University Press*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E.: 2007, 'Moses: Open Source Toolkit for Statistical Machine Translation'. In: *Proceedings of the Association for Computational Linguistics Companion Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic.

Lapata, M.: 2003, 'Probabilistic Text Structuring: Experiments with Sentence Ordering'. *Computational Linguistics* **29**(2), pp. 263–317.

Lapata, M.: 2006, 'Automatic Evaluation of Information Ordering: Kendall's Tau'. *Computational Linguistics* **32**(4), pp. 471–484.

Lavie, A. and A. Agarwal: 2007, 'METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments'. In: *Proceedings of the Workshop on Statistical Machine Translation at the Meeting of the Association for Computational Linguistics (ACL-2007)*. pp. 228–231.

Liang, Percy and Taskar, Ben and Klein, Dan: 2006, 'Alignment by Agreement'. In: *Proceedings of the Human Language Technology Conference of NAAC*. pp. 104–111.

Lin, C.-Y. and F. Och: 2004, 'Orange: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation.'. In: *Proceedings of the conference on Computational Linguistics*. p. 501.

Padó, S., M. Galley, C. D. Manning, and D. Jurafsky: 2009, 'Textual Entailment Features for Machine Translation Evaluation'. In: *the EACL Workshop on Machine Translation (WMT)*.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu: 2002, 'BLEU: a Method for Automatic Evaluation of Machine Translation'. In: *Proceedings of the Association for Computational Linguistics*. Philadelphia, USA, pp. 311–318.

Ronald, S.: 1998, 'More Distance Functions for Order-Based Encodings'. In: *the IEEE Conference on Evolutionary Computation*. pp. 558–563.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul: 2006, 'A Study of Translation Edit Rate with Targeted Human Annotation'. In: *AMTA*.

Ulam, S.: 1972, 'Some Ideas and Prospects in Biomathematics'. In: *Annual Review of Biophysics and Bioengineering*, pages 277–292.
Hirschberg, D.: 1975, 'A Linear Space Algorithm for Computing Maximal Common Subsequences'. In: *Communications of the ACM*, pages 341–343.

Address for Offprints:
University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, UK