A Cluster-Based Representation for Multi-System MT Evaluation

Nicolas Stroppa, Karolina Owczarzak
School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland
{nstroppa,owczarzak}@computing.dcu.ie

Abstract

Automatic evaluation metrics are often used to compare the quality of different systems. However, a small difference between the scores of two systems does not necessary reflect a real difference between their performance. Because such a difference can be significant or only due to chance, it is inadvisable to use a hard ranking to represent the evaluation of multiple systems.

In this paper, we propose a cluster-based representation for quality ranking of Machine Translation systems. A comparison of rankings produced by clustering based on automatic MT evaluation metrics with those based on human judgements shows that such interpretation of automatic metric scores provides dependable means of ordering MT systems with respect to their quality. We report experimental results comparing clusterings produced by BLEU, NIST, METEOR, and GTM with those derived from human judgement (of adequacy and fluency) on the IWSLT-2006 evaluation campaign data.

1 Introduction

Automatic evaluation metrics for Machine Translation (MT) have been given a lot of attention in the recent years, as their importance for MT research is hard to ignore. They are extremely useful in comparisons of developmental stages of an MT system, helping to test the influence of various parameters on the final translation output: addition or modification of rules in rule-based MT systems, modification of training settings for data-driven MT systems, etc. Moreover, they are also often used to compare the quality of different systems. Several evaluation campaigns strongly rely on automatic evaluation metrics (NIST, 2006; Paul, 2006) as well as on human judgment, which remains the ultimate evaluation schema, to assess the quality of participating MT systems.

The rankings of MT systems obtained with automatic evaluation metrics or human judgment are not strict in the sense that those scores may not be sufficient to distinguish between two systems. Indeed, a small difference between two scores does not necessary reflect a real difference between the performance of two systems. To test if the difference between the scores of two systems is *significant* or only due to chance, we can employ statistical significance tests using bootstrap (Efron and Tibshirani, 1993; Koehn, 2004) or approximate randomization (Noreen, 1989; Riezler and Maxwell, 2005) methods. This enables us to introduce a cluster-based representation which we feel is better suited to the ranking of system scores than a strict ranking which might be based on insignificant or accidental differences.

The quality of an automatic metric is often assessed by computing its correlation with human judgment (of adequacy and fluency) on a segment or system level. For an automatic evaluation metric, a high correlation with human judgment denotes a capability to correctly identify the quality of an MT system. In this paper, instead of com-

puting the direct correlation between automatic scores and human scores on a segment level, or in a hard ranking on a system level, we compare the clusters produced by automatic metrics and human judgements using an adaptation of the Rand statistic. In other terms, in this context, a metric will be considered good if it ranks various systems in the same order and groups them in the same clusters as human evaluators. We extend our analysis to clusterings produced by several automatic MT evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), and GTM (Melamed et al., 2004), using the evaluation data from the IWSLT-2006 evaluation campaign (Paul, 2006).

The remainder of this paper is organized as follows. In Section 2, we introduce the automatic evaluation metrics we tested. In Section 3, we present a statistical significance test based on approximate randomization, the cluster-based representation for ranking, and the cluster comparison strategies. In Section 4, we report experimental results. Section 5 concludes the paper and gives avenues for future work.

2 Automatic Evaluation Metrics

Since the introduction of BLEU (Papineni et al., 2002), a large number of other metrics have been developed, but the string-based metrics like BLEU, NIST (Doddington, 2002), GTM (Melamed et al., 2004), and METEOR (Banerjee and Lavie, 2005) have remained among the most popular, therefore we focus our analysis on them.

2.1 BLEU

The most popular evaluation metric BLEU (BiLingual Evaluation Understudy, (Papineni et al., 2002)) is based on a simple calculation of *modified precision*. Modified precision counts the number of n-grams in the translation that match at least one of the references and caps the count by the maximum number of occurrences of a given n-gram in a single reference. In other words, if a translation consists entirely of the word *the* repeated five times, but in one of the references *the* appears only once, and in the other only twice, we are allowed to count only two of the five matching words. This process is applied to any n, but in practice n-grams up to four are used. The modified precision results *for the whole document* at each n-gram level are combined together using geometric average. Moreover, in order to prevent unfair high precision scores for very short sentences, a brevity penalty is calculated over the test set, if the combined length of the translation segments is equal to or shorter than the combined length of best matching (closest in length) reference segments.

Note that BLEU was developed with document- or system-level evaluation in mind, and its construction does not allow for high correlation with human judgment on the level of individual segments. At segment level, many sentences will be scored as zero for not providing at least one four-gram in common with the references, which artificially levels down their quality. Segments shorter than four elements will be scored as zero irrespective of the number of lower n-gram matches. These effects are exacerbated as the number of available references decreases.

2.2 NIST

NIST was developed on the basis of BLEU-style n-gram calculation, but several improvements were added to raise the metric's correlations with human judgments (Doddington, 2002). Instead of geometric average, arithmetic average is used to combine results from all levels up to five grams, and the brevity penalty was adjusted to minimize the impact of small length variations. Most importantly, all n-grams are weighted according to their information with respect to the reference sentences, so that rarer and more informative sequences present in the translation will contribute more to the final score than sequences that are more common, and thus less informative.

2.3 GTM

Exploring a different avenue of research, GTM uses the standard notions of precision, recall, and their composite F-measure, to evaluate translation quality (Melamed et al., 2004). It calculates the word overlap between the translation and the reference(s), preventing double-counting when a word occurs multiple times, and it caps the resulting number of matches by the mean length of the references. While it also has the option of weight-

ing contiguous sequences more than unconnected matching fragments, Turian et al. (2003) conclude from their experiments that such a weight lowers the correlation with human judgment. In this work, we thus use the unweighted version of GTM. Turian et al. (2003) also show that GTM outperforms both BLEU and NIST with respect to correlation, irrespective of the number of references available.

2.4 METEOR

The evaluation in METEOR (Banerjee and Lavie, 2005) proceeds in several stages. First, all exact matches between the translation and the reference are found; next, the remaining words are stemmed and the matching process repeats; finally, there is the option of using WordNet to find matches between synonyms among the remaining non-matched words. The final score combines precision and heavily weighted recall at the unigram level with a penalty for non-contiguous matches.

3 Comparing Multiple Systems

3.1 Statistical Significance Testing using Approximate Randomization

Since a small difference between the scores of two systems does not necessary reflect a real difference between their performance, it is important to identify when this difference is *significant* or only due to chance. To discriminate between these two cases, we assume a null hypothesis which states that the two systems are of the same quality, and we consider the difference between their scores as significant only if we find statistical evidence indicating that the null hypothesis is false (with a certain degree of confidence).

When assumptions can be made about the probability distributions yielding the scores, it is possible to employ parametric methods such as the Student's t-test. When no specific assumption can be made, as it is the case for automatic evaluation metrics, we have to resort to non-parametric methods, such as bootstrap (Efron and Tibshirani, 1993; Koehn, 2004) or approximate randomization (Noreen, 1989; Riezler and Maxwell, 2005) methods. To use bootstrap, one would have to take the translation output of each MT system, produce a large number of samples from that output using sampling with replacement, and then create clusters of MT systems by collecting those with overlapping confidence intervals. However, in this paper we consider approximate randomization rather than bootstrap, following Riezler and Maxwell (2005) and Collins et al. (2005), who suggest that approximate randomization is more appropriate in such a context.

To compare the output of two systems using approximate randomization, we proceed as follows. First, we assume that we have access to n translations of the same sentences for the two systems. These translations are respectively denoted T (for system 1) and T' (for system 2), with $|T| = |T'| = n$. The set of reference translations for these sentences is denoted R. The score for T and T' are respectively $s = M(T, R)$ and $s' = M(T', R)$, where M denotes some metric (e.g. BLEU); their difference is $s - s'$.

Then, we build k new pairs of translation sets obtained by randomly permuting the translations in T and T', yielding the pairs $(T_1, T'_1), \ldots, (T_k, T'_k)$. For each $i \in 1..k$, the shuffle (T_i, T'_i) is obtained as follows: each pair of sentence in (T, T') is randomly shuffled with probability 0.5. Intuitively, if system 1 is better that system 2, then we obtain a lower score for the translations in T_i than for those in the original T, since T_i is obtained by replacing some translations in T with some translations from T' of lower quality. Consequently, in this scenario, we have $M(T_i, R) < M(T, R)$; similarly, we would also expect $M(T'_i, R) > M(T', R)$. In short, we expect the newly created T_i to be of lower quality than the original T.

$$M(T_i, R) - M(T'_i, R) < M(T, R) - M(T', R).$$

If this inequality is verified for $i \in 1..k$, we set $v_i = 0$, and $v_i = 1$ otherwise. If system 1 is better that system 2, then we expect $\sum_{i=1}^{k} v_i$ to be close to 0. On the contrary, if system 1 is not significantly better than system 2, then shuffling translations has little effect on the difference between the scores obtained, and $\sum_{i=1}^{k} v_i$ is unlikely to be close to 0. The p-value is simply computed as fol-

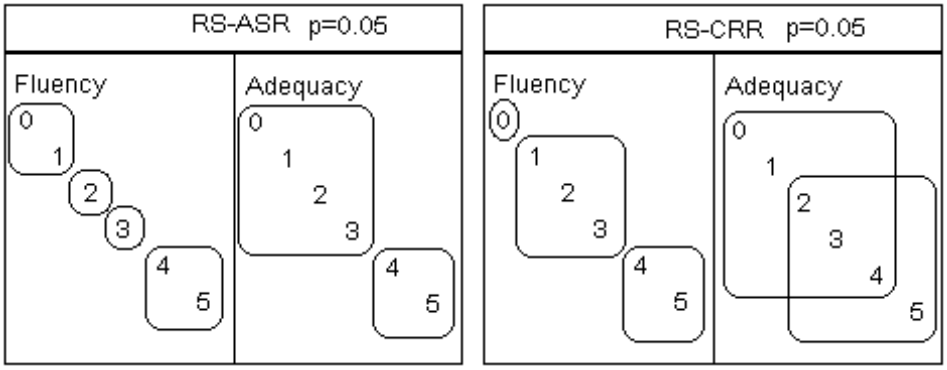

Figure 1: Examples of clusterings. Numbers 0-5 represent MT systems; clusters are created on the basis of fluency and adequacy scores. Relative height of the clusters shows their order.

lows:

$$p = \frac{(\sum_{i=1}^{k} v_i) + 1}{k}.$$

The null hypothesis is rejected if p is less than or equal to a specified rejection level, traditionally set to 0.05. In all our experiments, we used $k = 1000$ shuffles. We use the same method for all the considered metrics, including human judgement.

3.1.1 Implementation Issues

In order to compute statistical significance using approximate randomization, the values $M(T_i, R)$ and $M(T_i', R)$ are required for each shuffle (T_i, T_i'). However, even for document-level metrics such as BLEU, we do not have to compute BLEU for each shuffle. Indeed, it is sufficient to keep some information about each sentence (for BLEU: number of matching n-grams, lengths, etc.), and to aggregate them.

Consequently, the potentially expensive comparison between the reference sentences and the test sentences is performed once; only the aggregation of the sentence-level information, which is fast and cheap, is performed $k = 1000$ times. The computation of statistical significance for a test set of 500 sentences, with $k = 1000$ shuffles takes about 0.3 second for BLEU, and 0.7 second for NIST on a Pentium 4 processor, 3GHz.[1]

3.2 A Cluster-Based Representation

Most if not all comparisons of different MT sytems, including large-scale evaluations conducted in shared MT tasks, is done using a hard ranking of the participating sytems based on the system-level scores. However, as has been noted already, the difference in scores between two MT sytems may not be significant. We feel therefore that such strict rankings are inadvisable and not completely fair to the participating systems. In order to represent the ranking of MT systems according to their scores, we thus propose a cluster-based representation. In this representation, two systems are placed in the same cluster if they cannot be proven to differ in quality, i.e. if we have not succeeded in discarding the null hypothesis using approximate randomization. A cluster thus contains systems that are pairwise indistinguishable. By performing this comparison for all pairs of systems, this approach yields an ordered set of clusters. Formally, the method is expressed as follows. We note $s_1, s_2, \ldots, s_n$ the scores of n systems. We note $s_1 \gg s_2$ if s_1 is significantly higher than s_2, and $s_1 \sim s_2$ if their difference is not statistically significant. Using this cluster-based representation, we obtain an ordered set of clusters $C_1, \ldots, C_m$, such that:

$$\forall i \in 1..m,\ \forall k, l \in C_i,\ s_k \sim s_l,$$

$$\forall i, j \in 1..m,\ s.t.\ i < j,$$
$$\exists k \in C_i, l \in C_j, s_k \gg s_l.$$

This representation is suited to the ranking of system scores, and differs from the initial hard ranking, because one system can belong to several clusters. By using different p-values, we may

[1] Our C++ implementation, called FastMtEval, can be freely downloaded from `http://www.computing.dcu.ie/~nstroppa/softs/fast_mt_eval.tgz`.

obtain different cluster-based representations: the smaller the p-value, the bigger the clusters. An example of such a representation is given in Figure 1.

3.3 Comparing Clusters

In this section, we introduce a simple method to compare two clusterings. Our method is actually a simple adaptation of the Rand statistics (cf. Halkidi et al. (2001)), a method that can be used to compare non-ordered clusterings. The adaptation we propose aims at dealing with the ordered nature of the clusterings we consider.

A clustering C of n systems is a ordered set of clusters $C = \{C_1, \ldots, C_m\}$ such that $\forall i \in 1..m$, $C_i \subseteq 1..n$, and $\cup_i^m C_i = 1..n$. Let us recall that a system may belong to several clusters, i.e. we do not have necessarily $C_i \cap C_j = \emptyset$ for $i \neq j$.

To compare two clusterings C and D, we rely on a pairwise comparison of systems, i.e. clusterings C and D will be considered similar if for all pairs (i, j) of systems, C and D agree on the fact that systems i and j should be put on the same cluster or not. The Rand statistics counts the number of such agreements and divides it by the total number of comparisons, i.e. $\frac{n\times(n-1)}{2}$. In the ordered case, we have to add another factor. Indeed, if C and D agree that i and j should be placed on different clusters, but C says that i is significantly better than j and D shows the opposite, there is a strong disagreement between the clusterings. For a clustering C, we note $C(i, j)$ the relationship between the systems i and j according to the clustering. We have $C(i, j) \in \{'\sim', '\ll', '\gg'\}$. We have $'\sim'$, $'\ll'$, and $'\gg'$ respectively when i and j are indistinguishable, when j is significantly better than i, and when i is significantly better than j. The scoring is as follows:

$$s(c, d) = \begin{cases} 1 & \text{if } (c = d) \\ -1 & \text{if } (c = '\ll') \text{ and } (d = '\gg') \\ -1 & \text{if } (d = '\ll') \text{ and } (c = '\gg') \\ 0 & \text{otherwise.} \end{cases}$$

The first case corresponds to an agreement, the second and third cases are strong disagreements, and the last one is a weak disagreement. Our comparison metric is then computed as follows:

$$S(C, D) = \frac{2 \times \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s(C(i, j), D(i, j))}{n \times (n - 1)},$$

which yields a value between -1 and 1. A value of -1 denotes a complete disagreement on the ranking, while a value of 1 denotes a complete agreement.

For example, the "similarity" between the two clusterings associated with fluency and adequacy on the left of Figure 1 is 0.67. Indeed, they agree on the following (10) pairs: $(0, 1)$, $(0, 4)$, $(0, 5)$, $(1, 4)$, $(1, 5)$, $(2, 4)$, $(2, 5)$, $(3, 4)$, $(3, 5)$, $(4, 5)$, and (weakly) disagree on the following pairs: $(0, 2)$, $(0, 3)$, $(1, 2)$, $(1, 3)$, $(2, 3)$, which gives a final score of $\frac{2\times10}{6\times5} = 0.67$.

4 Experimental Results

4.1 Data

The experiments were carried out using the Chinese–English datasets provided within the IWSLT 2006 evaluation campaign (Paul, 2006), extracted from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). This multilingual speech corpus contains sentences similar to those that are usually found in phrase-books for tourists going abroad. Three input conditions are considered: continuous speech (CS-ASR), read-speech ASR (RS-ASR), and read-speech CRR (RS-CRR). In the first condition, the sentences to translate correspond to natural continuous speech; in the second case, the sentences are read and the input to translate comes from an ASR (Automatic Speech Recognition) system; in the last condition, MT systems are given the correct recognition results. For each conditions, 6 systems are considered. Since the various conditions corresponds to different views of the same sentences, it is possible to "merge" all the conditions together, in order to compare a total of 18 different systems (referred to as Mixed Track). The outputs of all systems were evaluated with respect to both adequacy and fluency. Automatic evaluation is performed using BLEU, NIST, METEOR, and GTM-1, with 7 references.

4.2 Cluster-Based Rankings

For each input condition and each metric, we constructed cluster-based rankings to represent the

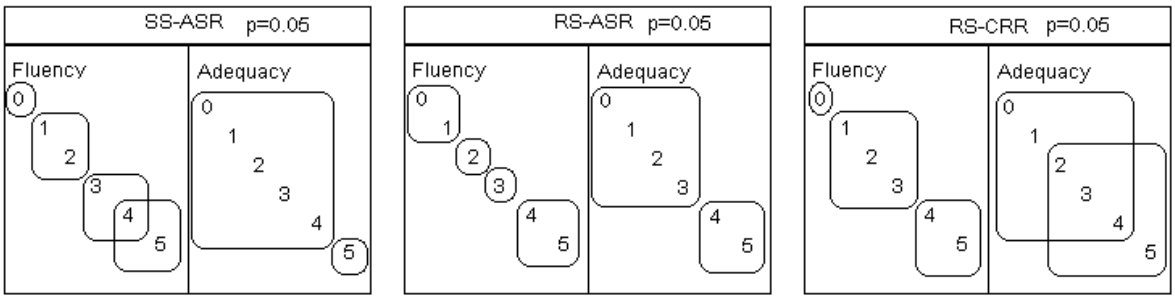

Figure 2: Clusterings of MT systems based on human judgements of fluency and adequacy

results obtained by the different systems. For those rankings, the level to test statistical significance is set to $p = 0.05$. The results for fluency and adequacy are displayed in Figure 2. Note that in this figure systems are numbered with respect to their rank according to a metric, i.e. system 0 in the fluency clustering is the best system according to fluency, and may be different from the system 0 in the adequacy clusterings.

We can observe that adequacy scores do not strongly differentiate the various participating systems, and the resulting clusters are big. In the case of fluency, there are more differences and systems are easier to distinguish. We also observe overlapping cases, in which a system belongs to several clusters.

To examine the influence of the significance level on the construction of the clusterings, we performed some tests with different values for p: 0.001, 0.002, 0.005, 0.01, 0.02, and 0.05. For the condition SS-ASR, we report the obtained results in Figure 3.

As expected, with a very high significance level ($p = 0.001$) it is not possible to distinguish between systems, and they are all placed in the same cluster, with respect to fluency as well as adequacy. Overall, however, the clusterings seem pretty stable: there are very few modifications between the clusterings with the p values 0.002, 0.005, 0.01, 0.02, and 0.05. For fluency, they are actually identical for the values 0.005, 0.01, and 0.02. For adequacy, they are identical for the values 0.002, 0.005, and 0.01. (See also Section 4.4 for a discussion about the choice of a significance level.)

4.3 Clusterings Comparison

Once constructed, we can compare the clusterings obtained with different evaluation metrics, using the comparison strategy introduced in Section 3.3 (with a p-value of 0.05). In particular, we computed the comparison scores between the automatic evaluation metrics BLEU, NIST, GTM-1, and METEOR, and the human judgement for fluency and adequacy. The results are displayed in Table 1.

		Fluency	Adequacy
SS-ASR	BLEU	0.47	0.4
	NIST	0	0.6
	METEOR	0	0.53
	GTM	-0.13	0.6
RS-ASR	BLEU	0.47	0.33
	NIST	0.4	0.27
	METEOR	0.33	0.13
	GTM	0.2	0.2
RS-CRR	BLEU	0.73	0.47
	NIST	0.4	0.27
	METEOR	0.53	0.26
	GTM	0.33	0.33
Mixed Track	BLEU	**0.58**	0.70
	NIST	0.34	0.64
	METEOR	0.39	**0.71**
	GTM	0.31	0.70

Table 1: Clustering comparison scores

According to these comparison scores, BLEU and METEOR seem to be better than NIST and GTM at finding rankings similar to those obtained with human judgement. In particular, BLEU yields consistently higher correlations with hu-

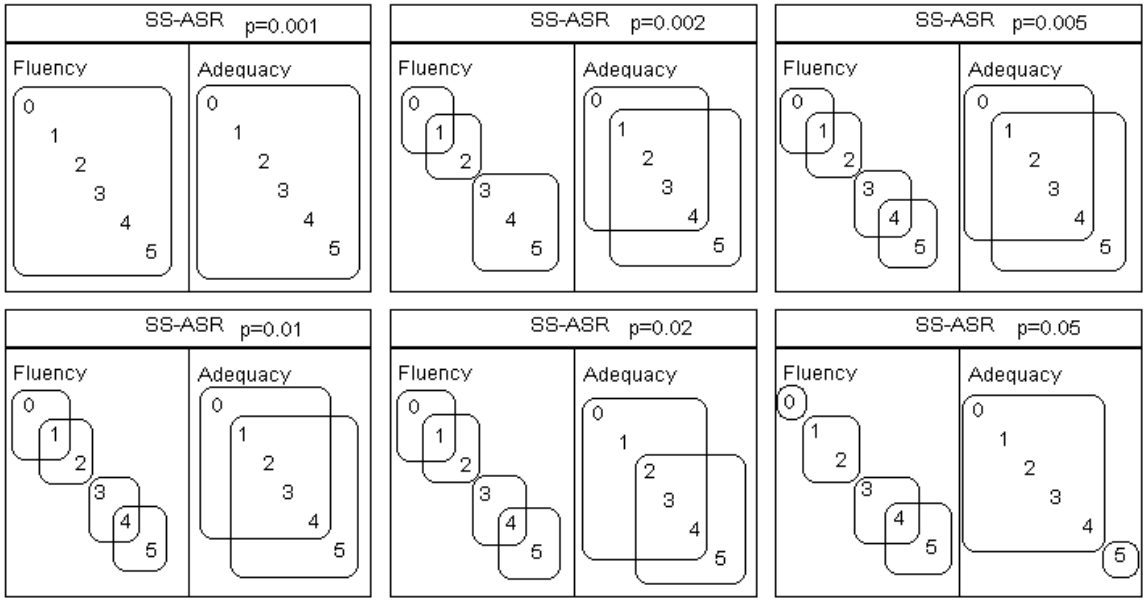

Figure 3: Clusterings obtained with diffrent p-values

man judgements of fluency, and GTM even obtains a negative score in the first input condition (spontaneous speech), showing a negative correlation with human ranking. In the case of adequacy, the picture is slightly less clear: BLEU seems to be more stable than the other metrics (it is better in two input conditions), even if METEOR has a higher correlation with adequacy in the Mixed Track. GTM-1 also achieves a high correlation for the Mixed Track. Let us also recall that this (indirect) approach based on the comparison of clusterings gives a view different from the computation of the direct correlation between segment-level or system-level hard rankings.[2]

We also compared how the clusterings obtained using the automatic evaluation metrics (BLEU, NIST, GTM-1, and METEOR) relate to each other. The results are displayed in Table 2.

Interestingly, the comparison scores between automatic evaluation metrics are higher than between the automatic evaluation metrics and the human judgement, which suggests that all these automatic metrics fall prey to some systematic error in evaluating translation quality.

[2]We do not claim that our method is better than direct correlation; instead it provides an alternative approach which is suited to the situation when an automatic metric is used to compare multiple systems.

	BLEU	NIST	METEOR
NIST	0.64	-	-
METEOR	0.77	0.79	-
GTM	0.70	0.79	0.86

Table 2: Comparing Automatic Metrics (Mixed Track)

4.4 Influence of the Significance Level

In Tables 1 and 2, the significance level is set to 0.05, since it is quite common to use such a value. However, this value affects the clusterings we obtain using our method (see e.g. Figure 3). In particular, a very small p-value (such as 0.001) yields inevitably a unique cluster containing all the systems, independently of the metric, which results in a correlation of 1 when comparing any two metrics. Obviously, there is a clear trade-off between the ability to produce a ranking and the level of confidence about this ranking.

In order to quantify the influence of this parameter, we compute the correlation between automatic and human evaluations, with various values of p. The results we obtain are displayed in Figure 4 for fluency and in Figure 5 for adequacy.

In terms of correlations with human judgements of fluency, the order of the automatic evaluation metrics does not seem to depend on signif-

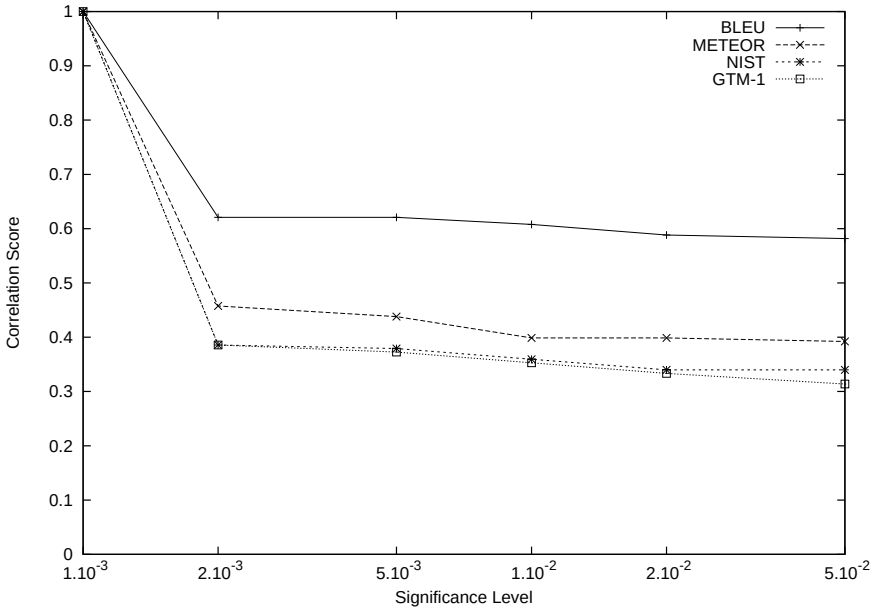

Figure 4: Influence of the p-value on the correlation with human judgements of fluency

icance level, and there is little variation between $p = 0.002$ and $p = 0.05$, although a very gentle decreasing trend can be noticed. Consequently, in this case, the choice of a p-value does not appear to be crucial. We can clearly observe that BLEU achieves the highest correlation with human judgements of fluency by a large margin.

Concerning adequacy, there is again little variation between $p = 0.002$ and $p = 0.05$, even if the relative order of the various metrics is not as stable. However, it seems that METEOR and GTM-1 are consistently better than the two other metrics, at least until $p = 0.05$.

5 Discussion and Conclusion

The variation in the number of clusters between tables in Figure 3 confirms the intuition that as the level of required confidence increases, it becomes more and more difficult to distinguish between different systems. The number of clusters ranges from one at $p = 0.001$, where all systems are seen as equal and the null hypothesis cannot be disproved, to four at $p = 0.05$ for fluency. Interestingly, clustering the systems with respect to their adequacy scores does not show the same level of refinement: at $p = 0.05$ there are only two (albeit non-overlapping) clusters. This tendency is not surprising, given that adequacy and fluency are two separate dimensions of a translation, each with its own set of conditions, so it is possible for systems to differ in the fluency of their output while being similar with respect to the semantic/lexical content. This duality of evaluation is often ignored in the creation of new automatic metrics for MT evaluation, where the guiding factor is usually the metric's correlation with the *average* human judgement.[3]

The comparison of clusters produced by BLEU, NIST, GTM, and METEOR on one hand, and human scores on the other, presented in Table 1, provides some surprising results. It turns out that BLEU, despite being widely criticised for low correlations with human judgements on segment level (Callison-Burch et al., 2006), consistently produces the most reliable clusters on the system level when it comes to judgements of fluency, and this trend is not influenced by the required significance level. Since BLEU was developed with system-level evaluation in mind, this is understandable; what is interesting, though, is that NIST, GTM, and METEOR, which are supposed to produce better segment-level evaluation than BLEU, are much worse than BLEU at

[3]Perhaps this is the reason why automatic metrics still seem so far away from successfully modeling human evaluation; it would be interesting to see whether we could devise a better metric by focusing on the two dimensions of fluency and adequacy separately.

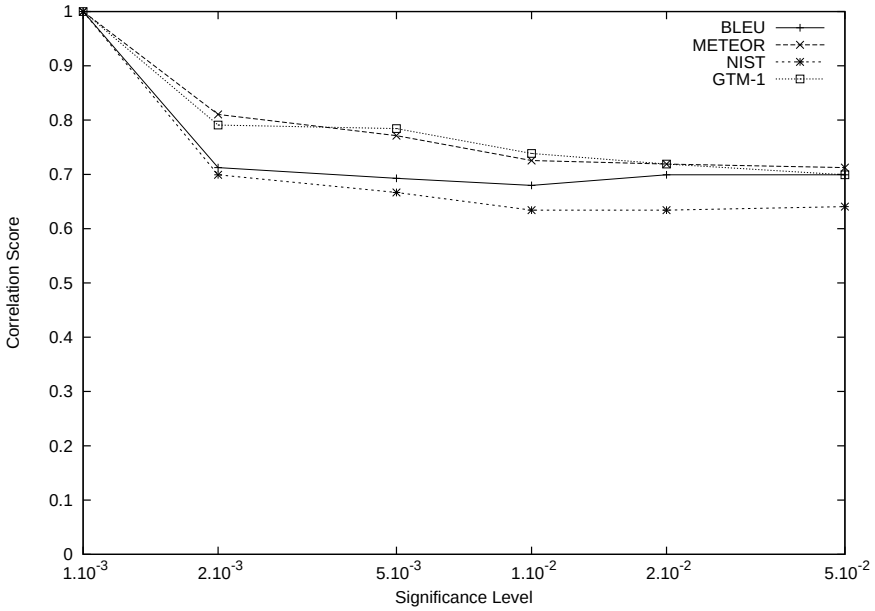

Figure 5: Influence of the p-value on the correlation with human judgements of adequacy

the system level - after all, we would expect the system-level evaluation to be directly dependent on the evaluation of its segments. This emphasizes the need to carefully choose one's metric depending on the type of task: it seems that for multiple system comparison BLEU does rather well, even though NIST, GTM, and METEOR might be more useful in the process of developing a single system (where the improvements often relate to specific types of sentences or structures and therefore a metric with a higher segment-level reliability would be better).

When it comes to correlations with human judgements of adequacy, these are on the whole higher for all the metrics; however, it must be remebered that the clusterings in the dimension of adequacy showed a much lower granularity than fluency, so it is easier to achieve high correlation. The difference between fluency and adequacy is smallest for BLEU, showing that a BLEU score reflects adequacy and fluency more equally than others. However, here BLEU is outperformed by METEOR and GTM-1, as the clusterings produced by these two metrics better reflect clusterings based on human judgement, at least for most values of p examined here. It seems then that here is where the advantage brought by better segment-level correlation with human judgement of METEOR and GTM is revealed.

Our future work includes conducting the clustering tests with a larger number of MT systems, to see whether the trends mentioned above hold in situations with a greater number of clusters. We also plan to add more metrics to our comparison, and vary the test with respect to the number of references available to the automatic metrics. Additionaly, we would like to compare the clusterings achieven in approximate randomization experiments with clusterings produced by a bootstrapping method for the same set of data.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the the ACL Workshop on Intrisic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL 2006*, pages 249–256, Trento, Italy.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL 2005*, pages 531–540, Ann Arbor, MI.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccur-

rence statistics. In *Proceedings of HLT 2002*, pages 128–132, San Diego, CA.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

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

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2004. Precision and recall of machine translation. In *Proceedings of HLT-NAACL 2003*, volume 2, pages 61–63, Edmonton, Canada.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience, New York, NY.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, PA.

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of IWSLT 2006*, pages 1–15, Kyoto, Japan.

Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the the ACL Workshop on Intrisic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 57–64, Ann Arbor, MI.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of LREC 2002*, pages 147–152, Las Palmas, Spain.