Truly Exploring Multiple References for Machine Translation Evaluation

Ying Qin
Dept. of Computer Science
Beijing Foreign Studies University
qinying@bfsu.edu.cn

Lucia Specia
Dept. of Computer Science
University of Sheffield
l.specia@sheffield.ac.uk

Abstract

Multiple references in machine translation evaluation are usually under-explored: they are ignored by alignment-based metrics and treated as bags of n-grams in string matching evaluation metrics, none of which take full advantage of the recurring information in these references. By exploring information on the n-gram distribution and on divergences in multiple references, we propose a method of n-gram weighting and implement it to generate new versions of the popular BLEU and NIST metrics. Our metrics are tested in two into-English machine translation datasets. They lead to a significant increase in Pearson's correlation with human fluency judgements at system-level evaluation. The new NIST metric also outperforms the standard NIST for document-level evaluation.

1 Introduction

Quality evaluation plays a critical role in Machine Translation (MT). Since its conception, the BLEU metric (Papineni et al., 2002) has had a significant impact on MT development. Although human evaluation has been used in recent evaluation campaigns such as WMT (Workshop on Statistical MT) (Bojar et al., 2014) and other forms of reference-less metrics have been proposed (Gamon et al., 2005; Specia et al., 2010), the merit of language and resource-independent n-gram based metrics such as BLEU is undeniable. Despite its criticisms, BLEU is thus still considered the *de facto* or at least a baseline metric for MT quality evaluation.

Due to the cost of human translation, often only one reference translation is available at evaluation time. However, generally there are numerous valid translations for a given sentence or document. Different references provide valid variations in linguistic aspects such as style, word choice and word order. Therefore, having multiple reference translations is key to improve the reliability of n-gram based evaluation metrics: the more references, the more chances for n-grams correctly translated to be captured. HyTER, an n-gram matching metric based on an exponential number of reference translations for a given target sentence, demonstrates the potential for better machine translation evaluation results from having as many references as possible (Dreyer and Marcu, 2012). Nevertheless, in the more realistic case where only a few references are available, if these are simply taken as bags of n-grams, increasing the number of references will not lead to the best possible results, as pointed out by Doddington (2002).

In this paper we explore how to use multiple references by means other than simply viewing them as bags of n-gram like BLEU, NIST (Doddington, 2002) and other n-gram co-occurrence based metrics do. Our assumption is that each reference reflects the complete meaning of the source segment. The semantic entirety of the translation will be adversely affected if all the n-grams from various references are simply put together. We propose a method of modifying the weight assignment strategy in BLEU and NIST by taking into account the n-gram distributions and divergences over different references.

Experiments were performed on two into-

© 2015 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

English translation datasets released by LDC, leading to promising results. In the remainder of this paper we will first review BLEU and related n-gram based evaluation metrics (Section 2). We then describe the method we propose to explore multiple references by reassigning the weights of n-grams that are common in system translations and references (Section 3), and the experiments performed and their results (Section 4). These illustrate how the modified BLEU and NIST scores compare against standard BLEU and NIST scores at the system, document and sentence levels.

2 N-gram based evaluation

2.1 BLEU

The BLEU metric applies a straightforward method of counting the n-grams that overlap in the system translation and given human translations under the assumption that human translations precisely reproduce the meaning of the source text. The closer to the reference, the higher the translation quality of the system translation will be. The core formula is given in Eq. 1 (Papineni et al., 2002), so that we can subsequently compare it to our approach.

$$S\_B = BP \times exp \sum_{n=1}^{N} w_n log P_n, \qquad (1)$$

where

$$P_n = \frac{\sum_{C \in Candi} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C \in Candi} \sum_{ngram' \in C'} Count(ngram')}$$

$$BP = \begin{cases} 1, if|c| \geq |r| \\ \mathrm{e}^{(1-|r|/|c|)}, if|c| < |r| \end{cases}$$

w_n is a weighting factor usually set as $1/N$, where N is the longest possible n-gram considered by the matching method. N is usually set to 4 to avoid data sparseness issues resulting from longer n-grams. P_n is the n-gram precision at a given n and in essence represents the proportion of n-grams in the candidate translation that also appear in the reference translation. BP is a penalty factor for shorter segments. c and r are the length of the candidate segment and reference segment, respectively.

When multiple references are available, $Count_{clip}(ngram)$ is clipped at the maximum count of n-grams which occurs in a single reference, and r is set as the length of the reference closest in size to the candidate translation.

Due to the sparsity of n-grams with large n and the geometric average of n-gram precisions, BLEU is not suitable for sentence-level evaluation. Several smoothing approaches have been proposed to alleviate this issue, such as the standard plus-one smoothing (Lin and Och, 2004) and combinations of smoothing techniques (Chen and Cherry, 2014).

A great deal of methods have been proposed to improve the performance of BLEU. These include metrics such as m-bleu (Agarwal and Lavie, 2008) and Amber (Chen and Kuhn, 2011). However, these metrics still treat n-grams in different references equally, regardless of whether the n-gram appears only once or is found in all references.

2.2 NIST

The NIST metric weights n-grams that occur less frequently in references more heavily (Doddington, 2002), as shown in Eq. 2.

$$S\_N = \sum_{n=1}^{N} \left\{ \sum_{\substack{w_1 \ldots w_n \\ co-occur}} Info(w_1 \ldots w_n) / \sum_{\substack{w_1 \ldots w_n \\ \text{in system}}} (1) \right\} \times exp\left\{ \beta log_2 \left[min(\frac{L_{sys}}{\overline{L_{ref}}}, 1) \right] \right\}, \qquad (2)$$

where

$$Info(w_1 \ldots w_n) = log(\frac{\text{\# of occur of } w_1 \ldots w_{n-1}}{\text{\# of occur of } w_1 \ldots w_n})$$

and $\overline{L_{ref}}$ is the average number of words in all references, L_{sys} is the number of words in the system translation, β is used as a weight for the penalty factor, and N is often 5.

The NIST metric focuses on non-popular n-grams in references and assumes that highly frequent n-grams, such as function words, tend to carry little meaning. However, this method consequently weakens the validity of n-grams that recur in multiple references. Since all references are valid translations for the same source text, one would expect multiple references to share common words and phrases that convey core meaning. Therefore, reducing the importance of these common n-grams is not beneficial to quality evaluation.

2.3 Improvements on n-gram based metrics

Current improvements on the n-gram co-occurrence evaluation metrics can be divided into three categories. The first category extends the scope of similarity detection by using a more

flexible matching strategy, for example using WordNet to capture synonyms as in METEOR (Banerjee and Lavie, 2005). The second category uses different functions to calculate the degree of similarity, for example edit distance, error rate, semantic distance (Nießen et al., 2000; Leusch et al., 2003; Snover et al., 2006; Snover et al., 2009). And the last category weights or combines the outcome of similarity functions as features (Liu et al., 2010; Giménez and Màrquez, 2010).

These methods focus on different forms of comparison between candidates and references. However, to our knowledge there are no other attempts to mine recurring information from multiple references if these are provided. Assuming all the possible translations form a "semantic" space, each reference only covers a subspace. The recurring n-grams among them should constitute the core part of this semantic space, which is more likely to represent the meaning of the source text. It is this kind of information that we want to explore and apply with our n-gram weighting technique.

3 Exploring information from multiple references

Although references can vary with translators and styles, many essential words and expressions are usually expected to be identical or similar for the same source text. For example, consider the segments below from our datasets: four references and one system (Sys) translation.

Ref1: *The gunman was shot to death by the police.*

Ref2: *Police killed the gunman.*

Ref3: *The gunman was shot dead by the police.*

Ref4: *The gunman was shot to death by the police.*

Sys: *Gunman is shot dead by police.*

Four unigrams appear in all four references: *., the, gunman, police*. The words *shot* and *by* appear three times whilst *dead* only appears once. The most recurring content unigrams in the references convey most of the meaning of the sentence. For the system output, there are six unigrams matching those in references, among them *gunman, police*, which occur in all references. However these are equally counted by BLEU and set as to have the lowest information value by NIST, compared to other unigrams such as *dead*, which only occurs once in one reference. This results in very low scores. The smoothed BLEU score for this segment is 0.3217 since there are no 3/4-grams matchings. The NIST score is 2.8867. However this is a rather good translation, with human judgements on fluency and accuracy of 4 and 4.7, respectively (human judgement ranges over 1-5). Taking into account the recurring n-grams in multiple references and assigning them heavier weights could thus be helpful to capture the quality of this system translation.

If function words are disregarded, an n-gram that recurs in most of the references could represent the core meaning of the source. The more often an n-gram is found in multiple references, the higher the probability that a matching n-gram appears in a high quality translation. Therefore, focusing on common n-grams found in multiple references, we propose a modified n-gram weighting approach for BLEU and NIST on the basis of the following factors.

3.1 Frequency of recurring n-grams in references

The degree of n-gram recurrence among references is represented by the number of times an n-gram appears in the references M divided by the total number of references $refno$. Nevertheless, it is unlikely that the number of times an n-gram occurs in references increases the significance of the respective n-gram by that number compared to an n-gram that occurs only once. Therefore we use the logarithm ratio instead. As an n-gram may be contained in all references, the add-one approach is then applied to avoid the expression in the logarithm returning a value of zero, as in Eq. 3.

$$log(1 + M/refno) \tag{3}$$

This attempt to reweight n-grams in BLEU and information content in NIST however did not lead to satisfactory results. Upon further analysis of the weighting strategy, we discovered that it is biased towards n-grams with a small n whose co-occurrence probability may be much higher than n-grams with a large n. In other words, the weighting is biased towards high-frequent function words, thus deviating from our original intention of assigning heavier weight to content (recurring) n-grams. As a result, using frequency as the only factor for n-gram reweighting is insufficient to capture useful information in multiple references.

3.2 Divergence of n-grams

In order to reduce the weight of most frequent function words, the distribution of n-grams is taken into account to improve Eq. 3. Less overlap among references may indicate that the translation is difficult, or that several different valid translations exist. In this scenario, recurring n-grams tend to be function words rather than content words. For instance, only function words repeat in the three references below, which may indicate that the source can be translated in many ways:

a. *At this time, the police have blocked the bombing scene.*

b. *They have now sealed off the spot.*

c. *The police has already blockaded the scene of the explosion.*

To address the problem, a unit called n-gram divergence is defined as in Eq. 4 to describe the degree of concentration of n-grams among references. The more divergent the distribution of n-grams in the references, the lower weight that is assigned to the most frequent common n-grams in the references.

$$Ngram_{diver} = \frac{\text{\# type of n-gram}}{\text{\# total of n-gram}}, \qquad (4)$$

i.e. the count of different n-grams divided by total number of n-grams. The higher the number of n-gram types found in multiple references, the more flexible or variable the translation will be, resulting in a higher value for n-gram divergence. This unit is used to measure the degree to which multiple references are similar.

3.3 Length of n-grams

The quality of the translation improves with the length of the matching n-grams, both in terms of fluency and accuracy evaluation. An additional modification of Eq. 3 is performed by replacing the constant 1 with the length of n-gram n, as depicted in Eq. 5,

$$log(n + M/refno). \qquad (5)$$

Eq. 6, denoted as R, is the final expression applied to reweight n-grams in BLEU and NIST and incorporates all of the factors described above.

$$R = Ngram_{diver} \times log(n + M/refno) \qquad (6)$$

3.4 Using Zipf's law

An alternative approach of neutralising function words in references is to use the Zipf's law. Ha et al. (2002) verify Zipf's law on n-grams by ranking all n-grams ($n \geq 1$). So the n-grams recurring in references in Eq. 3 can be represented by the product between frequency f and the ranking order r of n-grams divided by $refno$, as in Eq. 7.

$$R' = log(1 + r \times f/refno) \qquad (7)$$

The new BLEU score, denoted as $S_{\_BM}$, i.e., Score of BM, is rewritten in Eq. 8,

$$S_{\_BM} = BP \times exp(\sum_{n=1}^{N} w_n log(R \times P_n)), \qquad (8)$$

where BP, w_n and P_n are as stated as in Eq. 1. Add-one smoothing is applied to the segment level evaluation. In the equation, R can be replaced by R'. We compare the performance of the two weighting approaches in our experiments.

The modified NIST score formula, denoted as $S_{\_NM}$ (Score of metric NM), is shown as Eq. 9.

$$S_{\_NM} = \sum_{n=1}^{N} \left\{ \sum_{\substack{w_1 \ldots w_n \\ co-occur}} Info(w_1 \ldots w_n) / \sum_{\substack{w_1 \ldots w_n \\ \text{in system}}} (1) \right\} \times R \times exp\left\{ \beta log_2 \left[min(\frac{L_{sys}}{L_{ref}}, 1) \right] \right\} \qquad (9)$$

3.5 Arithmetic mean BLEU

Another modification in NIST with respect to BLEU is the fact that it uses arithmetic instead of geometric mean (Doddington, 2002). Although our method focuses on scenarios with multiple references in evaluation, further comparison to NIST is made by changing the averaging strategy in BM to that of NIST, denoted as BMA (BLEU Multi-reference Arithmetic mean).

4 Experiments and results

4.1 Data

Despite of the shortage of multiple references for MT evaluation, two datasets are found suitable to conduct experiments to test our reweighting strategy. The first dataset is Multiple-Translation Chinese Part 2 (MTC-P2) (LDC2003T17), including 4 sets of human translations for a single set of Mandarin Chinese source materials, 100 stories with 212-707 Chinese characters, totally 878 segments. There are three system translations P2-05,

P2-09 and P2-14 with human judgements on fluency and accuracy respectively. The other dataset is Multiple-Translation Chinese Part 4 (MTC-P4) (LDC2006T04), also with 4 references, 100 news stories each with 280-605 characters, totally 919 segments. Six system translations P4-09, P4-11, P4-12, P4-14, P4-15 and P4-22 are judged by 2-3 human annotators.

Human judgements for the nine system translations were carried out at segment level within limited time. Hence we firstly check the agreement among human annotators. We considered an agreement when two out of two judgements or two out of three judgements are same. The agreement proportion at system level is the number of segments agreed upon divided by the total number of segments in the system. This agreement proportion is normalised by the degree of agreement by chance, i.e., using Cohen's kappa coefficient which is commonly applied in WMT. Since the scale of human annotation is 1 to 5, the agreement by chance value is set as 0.2. Table 1 shows the kappa agreement of human annotators on all system translations. Note that the average agreement on fluency is only fair, while the agreement on accuracy is even worse. Given the subjectivity of the task, however, this range of figures is not uncommon.

	Flu	**Acc**
p2-05	0.311	0.254
p2-09	0.320	0.257
p2-14	0.294	0.280
p4-09	0.132	0.123
p4-11	0.143	0.094
p4-12	0.218	0.053
p4-14	0.247	0.106
p4-15	0.150	0.120
p4-22	0.229	0.264
Mean	0.227	0.172

Table 1: Kappa agreement of human judgement on system translations

Our evaluation is performed at the system, document and segment levels. Different human judgements are averaged for the final score of a segment, and all segment scores in a text are averaged for the final document score. While scores for smoothed BLEU and standard BLEU are similar at system and document levels, the standard BLEU score is generally below the smoothed BLEU score for segment level. BM is derived from smoothed BLEU.

4.2 System level

We compare the Pearson correlation for various automatic evaluation scores with human scores at system level in terms of fluency (Flu) and accuracy (Acc), as shown in Table 2.

	BLEU	**BM**	**BMA**	**NIST**	**NM**
Flu	0.7021	0.7090	**0.7136**	0.5657	0.5938
Acc	0.6957	0.6947	0.7114	**0.7941**	0.7756

Table 2: Pearson correlation at system level

For fluency judgements, BMA displays the highest correlation with human scores, 26.14% higher than NIST score and 1.64% better than BLEU. These results are promising. Compared to BLEU, BM is slightly better. NM scores also outperform NIST. The results are not as positive when measuring correlation to accuracy judgements. NIST still performs the best, however, the gap between BMA and NIST is much lower for accuracy than for fluency.

When we apply Eq. 7 to reweight BLEU, the correlation with human scores at system level achieves 0.6926 on fluency and 0.7391 on accuracy. This represents a distinct increase in correlation for accuracy judgements, making the gap to the best performing metric (NIST) even smaller. However, it leads to a slight decrease in correlation for fluency evaluation. Overall, our results demonstrate that the proposed methods is effective for fluency evaluation at system level.

4.3 Document level

Tables 3 and 4 shows the metrics comparison for document level evaluation. For fluency (Table 3), BM outperforms BLEU in 6 out of 9 systems, and its average correlation exceeds that of standard BLEU. BMA leads to even more promising results compared to BLEU. However at document level the BMA metric does not perform as well as NIST even using the same averaging method. Note that the performance of NM is better than standard NIST, indicating that the use of recurring n-grams in multiple references works. In fact, NM leads to the best fluency evaluation for all systems.

For accuracy evaluation (Table 4), the performance of BM, BMA and NM varies with different system outputs. NM still performs the best overall.

The reweighting approach in Eq. 7 is clearly inferior to BM at document level, with only 2 out of 9 outputs sightly better than BM both on fluency and accuracy evaluation. We speculate that

this may be because Zipf's law is less applicable to small scale datasets such as ours. Nevertheless, the n-gram weighting approach proposed in Eq. 6 proved effective.

	BLEU	BM	BMA	NIST	NM
p2-05	0.1510	0.1637	0.1627	**0.2495**	0.2401
p2-09	0.0990	0.0867	**0.0992**	0.0467	0.0653
p2-14	0.1666	0.1707	0.2102	0.2644	**0.2474**
p4-09	0.3423	0.3392	0.3716	**0.4343**	0.4291
p4-11	0.1310	0.1423	0.1492	0.1486	**0.1681**
p4-12	0.1479	0.1424	0.1711	0.1955	**0.2032**
p4-14	0.1168	0.1191	0.1373	**0.1610**	0.1577
p4-15	0.2384	0.2397	0.2703	**0.3189**	0.3163
p4-22	0.1568	0.1589	0.1660	0.2202	**0.2211**
Mean	0.1722	0.1736	0.1931	0.2266	**0.2276**

Table 3: Doc-level Pearson correlation on fluency

	BLEU	BM	BMA	NIST	NM
p2-05	0.2571	0.2621	0.2778	0.3334	**0.3549**
p2-09	0.0942	0.0874	**0.1015**	0.0936	0.0850
p2-14	0.2613	0.2633	0.2943	**0.3161**	0.3015
p4-09	0.3867	0.3808	0.4186	**0.4928**	0.4844
p4-11	0.1656	0.1825	0.1890	0.1604	**0.2016**
p4-12	0.3218	0.3197	0.3537	0.3751	**0.3847**
p4-14	0.1532	0.1495	0.1719	**0.1934**	0.1828
p4-15	0.2367	0.2292	0.2730	**0.4010**	0.3887
p4-22	0.0887	0.0922	0.0829	**0.2428**	0.2363
Mean	0.2184	0.2185	0.2403	0.2898	**0.2911**

Table 4: Doc-level Pearson correlation on accuracy

4.4 Segment level

In all datasets, BM performs worse at segment level than smoothed BLEU. The average gap in correlation between BM and BLEU is 4.5% on fluency and 2.9% on accuracy. NM outperforms NIST at segment level on 4 out of 9 systems on fluency, but overall, NM is slightly worse than NIST, for both fluency and accuracy.

We believe the main reason is that data sparsity of recurring n-grams at segment level is more severe than at document and system levels. The second possible cause is that the smoothed BLEU score is not based on actual n-gram matching between the candidates and references, but a predictable score computed even if there is no n-gram matching. It is hard to apply common information in multiple references to this score. Closer investigation is presented in the following section. Also important, the low agreement among humans on quality judgements might pose more challenges to evaluation than the methods themselves.

4.5 Discussion

Fluency and accuracy evaluation At system and document level, the reweighting strategy by considering multiple references yields better results than both BLEU and NIST. The improvements on fluency are much promising than on accuracy.

We examine the recurring n-grams in the four references in MTC-P2 in detail. Taking unigrams as example, among the unigrams in all references, 48.7% occur in a single reference, 17.8% are covered by any two references. As expected, the percentage of common n-grams decreases as we increase the number of references. There is a sharp drop when the number of references changes from one to two, indicating that most n-grams appear only in one reference. This becomes a more severe limitation of the dataset for n-grams with larger n, as depicted in Figure 1. 91.86% of 4-grams appear in a single reference, while only 0.24% are covered by the four references.

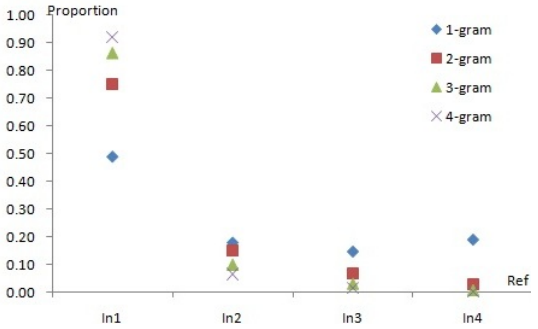

Figure 1: Common 1-4grams in references of MTC P2 (InX denotes covered by X references)

For the matching n-grams between a candidate and references, all n-gram counts but unigram counts go down as we increase the number of references. Figure 2 illustrates the distribution of matching n-grams for the P2-05 system as an instance. Among the matching unigrams, 20% appear in one of the references, 17% appear in two of them, 22% in three, and 41% are covered by all references. Notice that the matching unigrams that occur in all four references exceed the unigrams that appear in less than four references. However, most of these unigrams are function words and punctuation. Weighting them more heavily has a negative effect on accuracy evaluation, especially at segment level. This also explains the increase in correlation for accuracy when Zipf's law is applied to deduce the effect of function words. On

the other hand, many higher-order n-grams were found in more than one reference, which explains the improvements on fluency evaluation.

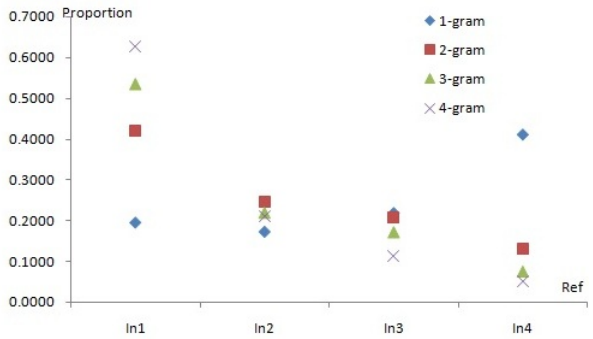

Figure 2: Matching n-grams distribution of P2-05

Content vs functional n-grams Using Eq. 7 to assign heavier weight to content n-grams improves the correlation for accuracy evaluation at system level, but leads to a drop in correlation for document and segment level evaluations. Thus there is no clear advantage for using such an approach to weighting function words and content n-grams differently, at least for the datasets used in the experiments.

Influence of number of references Our experiments use four references, only a small portion of the valid translations for the source texts. This somewhat limited the exploration of the proposed reweighting method.

Eq. 5 indicates that the larger the number of references, the lower the weighting ratio for recurring n-grams. For instance, for bigrams appearing twice in 10 references, the outcome of Eq. 5 is 0.3424, while for bigrams appearing once, the outcome of Eq. 5 is 0.3222. However since there are only four references, the weighting ratio is larger, 0.3979/0.3522. In other words, the larger the number of references, the lower the impact of the reweighting method on the results.

Increasing the number of references could help discriminate function words and content words as well. To check the recurrence of n-grams in larger numbers of references, we investigate the devset1-3 of BTEC (Takezawa et al., 2002), which contains 1512 source sentences, each with 16 English references. We show the average 1-4grams distribution over 2 to 16 translations in Figure 3. As expected, the proportion of n-grams covered by multiple references decreases as the number of references increases, showing that more translation variety is obtained with more references. The total number of 1-4grams found in three references (In3) is still as high as 28.4%, demonstrating the potential benefits of exploring multiple references.

5 Conclusions and future work

Recurring n-grams in references can help capture important words and sequences of words that are chosen by various translators. By combining recurrence distributions, divergence information and the length of n-grams, a modified weighting strategy for BLEU and NIST was proposed to make better use of multiple references in translation evaluation. This strategy was tested with different reweighting schemes. The results on two datasets proved promising.

Overall, the strategy favours fluency evaluation over accuracy evaluation. To address that, in future work we will further improve the metric by tackling common n-grams carrying lower information content. We also observed how the weaknesses of exact n-gram matching affects the performance of the proposed metrics. In future work, in addition to the n-gram distributions, divergence information and length of n-grams, synonym recurrence information will also be explored. Adapting this approach to other metrics such as METEOR is another direction for future work.

Acknowledgements

Ying Qin's work is supported by a Beijing Social Science Funding Project (15WYA006) and the National Research Centre for Foreign Language Education, BFSU.

References

Agarwal, A and A Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation. ACL*, pages 115–118.

Banerjee, S and A Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Bojar, Ondrej, Christian Buck, Christian Federmann, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58.

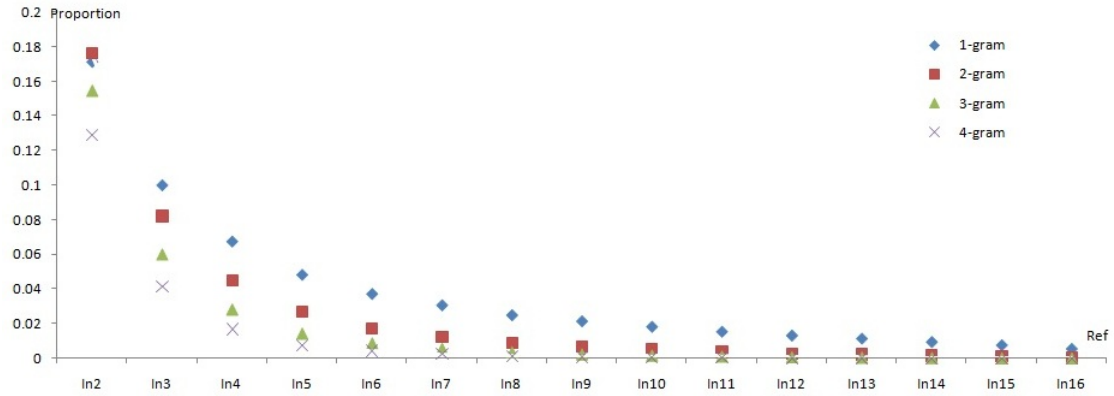

Figure 3: 1-4grams distribution in the BTEC corpus

Chen, Boxing and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *ACL 2014*, page 362.

Chen, Boxing and Roland Kuhn. 2011. Amber: A modified bleu, enhanced ranking metric. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 71–77.

Doddington, G. 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.

Dreyer, Markus and Daniel Marcu. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *2012 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 162–171.

Gamon, Michael, Anthony Aue, and Martine Smets. 2005. Sentence-level mt evaluation without reference translations: Beyond language modeling. In *Proceedings of EAMT*, pages 103–111.

Giménez, Jesús and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Ha, Le Quan, Elvira I Sicilia-Garcia, Ji Ming, and F Jack Smith. 2002. Extension of zipf's law to words and phrases. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–6. ACL.

Leusch, G, N Ueffing, and H Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*, pages 33–40.

Lin, Chin Yew and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics.ACL*, pages 501–507.

Liu, C, D Dahlmeier, and H. T. Ng. 2010. Tesla: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. ACL*, pages 354–359.

Nießen, Sonja, Franz Josef Och, Gregor Leusch, Hermann Ney, et al. 2000. An evaluation tool for machine translation: Fast evaluation for mt research. In *LREC*.

Papineni, K, S Roukos, T Ward, et al. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on ACL*, pages 311–318.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Snover, Matthew, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. ACL.

Specia, L, D Raj, and M. Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Takezawa, Toshiyuki, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *LREC*, pages 147–152.