Parmesan: Meteor without Paraphrases with Paraphrased References

Petra Barančíková
Institute of Formal and Applied Linguistics
Charles University in Prague, Faculty of Mathematics and Physics
Malostranské náměstí 25, Prague, Czech Republic
barancikova@ufal.mff.cuni.cz

Abstract

This paper describes Parmesan, our submission to the 2014 Workshop on Statistical Machine Translation (WMT) metrics task for evaluation English-to-Czech translation. We show that the Czech Meteor Paraphrase tables are so noisy that they actually can harm the performance of the metric. However, they can be very useful after extensive filtering in targeted paraphrasing of Czech reference sentences prior to the evaluation. Parmesan first performs targeted paraphrasing of reference sentences, then it computes the Meteor score using only the exact match on these new reference sentences. It shows significantly higher correlation with human judgment than Meteor on the WMT12 and WMT13 data.

1 Introduction

The metric for automatic evaluation of machine translation (MT) Meteor[1] (Denkowski and Lavie, 2011) has shown high correlation with human judgment since its appearance. It outperforms traditional metrics like BLEU (Papineni et al., 2002) or NIST (Doddington, 2002) as it explicitly addresses their weaknesses – it takes into account recall, distinguishes between functional and content words, allows language-specific tuning of parameters and many others.

Another important advantage of Meteor is that it supports not only exact word matches between a hypothesis and its corresponding reference sentence, but also matches on the level of stems, synonyms and paraphrases. The Meteor Paraphrase tables (Denkowski and Lavie, 2010) were created automatically using the *pivot* method (Bannard and Callison-Burch, 2005) for six languages.

[1]We use the the version 1.4., which was recently outdated as the new version 1.5. was released for WMT14

The basic setting of Meteor for evaluation of Czech sentences offers two levels of matches - exact and paraphrase. In this paper, we show the impact of the quality of paraphrases on the performance of Meteor. We demonstrate that the Czech Meteor Paraphrase tables are full of noise and their addition to the metric worsens its correlation with human judgment. However, they can be very useful (after extensive filtering) in creating new reference sentences by targeted paraphrasing.

Parmesan[2] starts with a simple greedy algorithm for substitution of synonymous words from a hypothesis in its corresponding reference sentence. Further, we apply Depfix (Rosa et al., 2012) to fix grammar errors that might arise by the substitutions.

Our method is independent of the evaluation metric used. In this paper, we use Meteor for its consistently high correlation with human judgment and we attempt to tune it further by modifying its paraphrase tables. We show that reducing the size of the Meteor Paraphrase tables is very beneficial. On the WMT12 and WMT13 data, the Meteor scores computed using only the exact match on our new references significantly outperform Meteor with both exact and paraphrase match on original references. However, this result was not confirmed by this year's data.

We perform our experiments on English-to-Czech translations, but the method is largely language independent.

2 Related Work

Our paraphrasing work is inspired by Kauchak and Barzilay (2006). They are trying to improve the accuracy of MT evaluation of Chinese-to-English translation by targeted paraphrasing, i.e. making a reference closer in wording to a hypothesis (MT output) while keeping its meaning and correctness.

[2]PARaphrasing for MEteor SANs paraphrases

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 355–361,
Baltimore, Maryland USA, June 26–27, 2014.

Having a hypothesis $H = h_1, ..., h_n$ and its corresponding reference translation $R = r_1, ..., r_m$, they select a set of candidates $C = \{\langle r_i, h_j \rangle | r_i \in R \setminus H, h_j \in H \setminus R\}$. C is reduced to pairs of words appearing in the same WordNet (Miller, 1995) synset only. For every pair $\langle r_i, h_j \rangle \in C$, h_j is evaluated in the context $r_1, ..., r_{i-1}, \square, r_{i+1}, ..., r_m$ and if confirmed, the new reference sentence $r_1, ..., r_{i-1}, h_j, r_{i+1}, ..., r_m$ is created. This way, several reference sentences might be created, all with a single changed word with respect to the original one.

In Barančíková et al. (2014), we experiment with several methods of paraphrasing of Czech sentences and filtering the Czech Meteor tables. We show that the amount of noise in the multi-word paraphrases is very high and no automatic filtering method we used outperforms omitting them completely. We present an error analysis based method of filtering paraphrases consisting of pairs of single words, which is used in subsection 3.1. From several methods of paraphrasing, we achieved the best results a with simple greedy method, which is presented in section 4.

3 Data

We perform our experiments on data sets from the English-to-Czech translation task of WMT12 (Callison-Burch et al., 2012), WMT13 (Bojar et al., 2013) and WMT14 (Bojar et al., 2014). The data sets contain 13/14[3]/10 files with Czech outputs of MT systems. In addition, each data set contains one file with corresponding reference sentences and one with original English source sentences. We perform morphological analysis and tagging of the hypotheses and the reference sentences using Morče (Spoustová et al., 2007).

The human judgment of hypotheses is available as a relative ranking of performance of five systems for a sentence. We calculated the score for every system by the "> others" method (Bojar et al., 2011), which was the WMT12 official system score. It is computed as $\frac{wins}{wins+loses}$. We refer to this interpretation of human judgment as *silver standard* to distinguish it from the official system scores, which were computed differently each year (here referred to as *gold standard*).

[3]We use only 12 of them because two of them (FDA.2878 and online-G) have no human judgments.

	WMT12	WMT13	WMT14
WordNet	0.26	0.22	0.24
filtered Meteor	1.53	1.29	1.39
together	1.59	1.34	1.44

Table 1: Average number of one-word paraphrases per sentence found in WordNet, filtered Meteor tables and their union over all systems.

3.1 Sources of Paraphrases

We use two available sources of Czech paraphrases – the Czech WordNet 1.9 PDT (Pala and Smrž, 2004) and the Meteor Paraphrase Tables (Denkowski and Lavie, 2010).

The Czech WordNet 1.9 PDT contains paraphrases of high quality, however, their amount is insufficient for our purposes. It contains 13k pairs of synonymous lemmas and only one paraphrase per four sentences on average is found in the data (see Table 1). For that reason, we employ the Czech Meteor Paraphrase tables, too. They are quite the opposite of Czech WordNet – they are large in size, but contain a lot of noise.

We attempt to reduce the noise in the Czech Meteor Paraphrase tables in the following way. We keep only pairs consisting of single words since we were not successful in reducing the noise effectively for the multi-word paraphrases (**?**).

Using Morče, we first perform morphological analysis of all one-word pairs and replace the word forms with their lemmas. We keep only pairs of different lemmas. Further, we dispose of pairs of words that differ in their parts of speech (POS) or contain an unknown word (typically a foreign word).

In this way we have reduced 684k paraphrases in the original Czech Meteor Paraphrase tables to only 32k pairs of lemmas. We refer to this table as filtered Meteor.

4 Creating New References

We create new references similarly to Kauchak and Barzilay (2006). Let H_L, R_L be sets of lemmas from a hypothesis and a corresponding reference sentence, respectively. Then we select candidates for paraphrasing in the following way: $C_L = \{(r, h) | r \in R_L \setminus H_L, h \in H_L \setminus R_L, r_{POS} = h_{POS}\}$, where r_{POS} and h_{POS} denote the part of speech of the respective lemma.

Further, we restrict the set C_L to pairs appearing in our paraphrase tables only. If a word has several

Source	*The location alone is classic.*
Hypothesis	*Samotné místo je klasické .* Actual place$_{neut}$ is classic$_{neut}$. The place alone is classic.
Reference	*Už poloha je klasická .* Already position$_{fem}$ is classic$_{fem}$. The position itself is classic.
Before Depfix	*Už místo je klasická .* Already place$_{neut}$ is classic$_{fem}$. *The place itself is classic.
New reference	*Už místo je klasické .* Already place$_{neut}$ is classic$_{neut}$. The place itself is classic.

Figure 1: Example of the targeted paraphrasing. The hypothesis is grammatically correct and has very similar meaning as the reference sentence. The new reference is closer in wording to the hypothesis, but the agreement between the noun and the adjective is broken. Depfix resolves the error and the final reference is correct. Number of overlapping unigrams increased from 2 to 4.

metric	reference	WMT12	WMT13
BLEU	original	0.751	0.835
	new	0.834	**0.891**
METEOR	original	0.833	0.817
	new	**0.927**	**0.891**
1 - TER	original	0.274	0.760
	new	0.283	0.781

Table 2: Pearson's correlation of different metrics with the silver standard.

paraphrases in C_L, we give preference to those found in WordNet or even better in both WordNet and filtered Meteor.

We proceed word by word from the beginning of the reference sentence to its end. If a lemma of a word appears as the first member of a pair in restricted C_L, it is replaced by the word form from hypothesis that has its lemma as the second element of that pair, i.e., by the paraphrase from the hypothesis. Otherwise, the original word the reference sentence is kept.

When integrating paraphrases to the reference sentence, it may happen that the sentence becomes ungrammatical, e.g., due to a broken agreement (see Figure 1). Therefore, we apply Depfix (Rosa et al., 2012) – a system for automatic correction of grammatical errors that appear often in English-to-Czech MT outputs.

Depfix analyses the input sentences using a range of natural language processing tools. It fixes errors using a set of linguistically-motivated rules and a statistical component it contains.

5 Choosing a metric

Our next step is choosing a metric that correlates well with human judgment. We experiment with three common metrics – BLEU, Meteor and TER. Based on the results (see Table 2), we decided to employ Meteor in WMT14 as our metric because it shows consistently highest correlations.

6 Meteor settings

Based on the positive impact of filtering Meteor Paraphrase Tables for targeted lexical paraphrasing of reference sentences (see the column **Basic** in Table 4), we experiment with the filtering them yet again, but this time as an inner part of the Meteor evaluation metric (i.e. for the paraphrase match).

We experiment with seven different settings that are presented in Table 3. All of them are created by reducing the original Meteor Paraphrase tables, except for the setting referred to as **WordNet** in the table. In this case, the paraphrase table is generated from one-word paraphrases in Czech WordNet to all their possible word forms found in CzEng (Bojar et al., 2012).

Prior paraphrasing reference sentences and using Meteor with the **No paraphr.** setting for computing scores constitutes Parmesan – our submission to the WMT14 for evaluation English-to-Czech translation. In the tables with results,

setting	size	description of the paraphrase table
Basic	684k	The original Meteor Paraphrase Tables
One-word	181k	**Basic** without multi-word pairs
Same POS	122k	**One-word** + only same part-of-speech pairs
Diff. Lemma	71k	**Same POS** + only forms of different lemma
Same Lemma	51k	**Same POS** + only forms of same lemma
No paraphr.	0	No paraphrase tables, i.e., exact match only
WordNet	202k	Paraphrase tables generated from Czech WordNet

Table 3: Different paraphrase tables for Meteor and their size (number of paraphrase pairs).

WMT12

reference	Basic	One-word	Same POS	Same Lemma	Diff. Lemma	No paraphr.	WordNet
Original	0.833	0.836	0.840	0.838	0.863	0.861	0.863
Before Depfix	0.905	0.908	0.911	0.911	0.931	0.931	0.931
New	0.927	0.930	0.931	0.932	0.950	**0.951**	**0.951**

WMT13

references	Basic	One-word	Same POS	Same Lemma	Diff. Lemma	No paraphr.	WordNet
Original	0.817	0.820	0.823	0.821	0.850	0.848	0.850
Before Depfix	0.865	0.867	0.869	0.868	0.895	0.895	0.894
New	0.891	0.892	0.893	0.892	**0.915**	**0.915**	**0.915**

Table 4: Pearson's correlation of Meteor and the silver standard.

Parmesan scores are highlighted by the box and the best scores are in bold.

7 Results

7.1 WMT12 and WMT13

The results of our experiments are presented in Table 4[4] as Pearson's correlation coefficient of the Meteor scores and the human judgment. The results in both tables are very consistent. There is a clear positive impact of the prior paraphrasing of the reference sentences and of applying Depfix. The results also show that independently of a reference sentence used, reducing the Meteor paraphrase tables in evaluation is always beneficial.

We use a freely available implementation[5] of Meng et al. (1992) to determine whether the difference in correlation coefficients is statistically significant. The tests show that Parmesan performs better than original Meteor with 99% certainty on the data from WMT12 and WMT13.

Diff. Lemma and **WordNet** settings give the best results on the original reference sentences. That is because they are basically a limited version of the paraphrase tables we use for creating our new references, which contain both all different lemmas of the same part of speech from Meteor Paraphrase tables and all lemmas from the WordNet.

The main reason of the worse performance of the metric when employing the Meteor Paraphrase tables is the noise. It is especially apparent for multi-word paraphrases (Barančíková et al., 2014); however, there are problems among one-word paraphrases as well. Significant amount of them are pairs of different word forms of a single lemma, which may award even completely non-grammatical sentences. This is reflected in the low correlation of the **Same Lemma** setting.

Even worse is the fact that the metric may award even parts of the hypothesis left untranslated, as the original Meteor Paraphrase tables contain English words and their Czech translations as paraphrases. There are for example pairs: *pšenice - wheat*[6], *vůdce - leader*, *vařit - cook*, *poloostrov - peninsula*. For these reasons, the differences among the systems are more blurred and the metric performs worse than without using the paraphrases.

[4]The results of WMT13 using the gold standard are in Table 5.

[5]http://www.cnts.ua.ac.be/∼vincent/scripts/rtest.py

[6]In all examples the Czech word is the correct translation of the English side.

	WMT13						
references	Basic	One-word	Same POS	Same Lemma	Diff. Lemma	No paraphr.	WordNet
Original	0.856	0.859	0.862	0.860	0.885	0.883	0.884
Before Depfix	0.894	0.896	0.898	0.897	0.918	0.917	0.917
New	0.918	0.918	0.919	0.919	**0.933**	**0.933**	**0.933**

Table 5: Pearson's correlation of Meteor and the gold standard – *Expected Wins* (Bojar et al., 2013). The results corresponds very well with the silver standard in Table 4.

	frequency	Basic	No paraphr.
WMT12	0.75	0.837	0.869
WMT13	0.61	0.818	0.852

Table 6: The *frequency* column shows average number of substitution per sentence using the original Meteor Paraphrase tables only. The rest shows Pearson's correlation with the silver standard using these paraphrases.

We also experimented with paraphrasing using the original Meteor Paraphrase tables for a comparison. We used the same pipeline as it is described in Section 4, but used only original one-word paraphrases from the Meteor Paraphrase tables. Even though the paraphrase tables are much larger than our filtered Meteor tables, the amount of substituted words is much smaller (see Table 6) due to not being lemmatized. The **Basic** setting in Table 6 corresponds well with the setting **One-word** in Table 4 on original reference sentences. The results for **No paraphr.** setting in Table 6 outperforms all correlations with original references but cannot compete with our new reference sentences created by the filtered Meteor and WordNet.

7.2 WMT14

The WMT14 data did not follow similar patterns as data from two previous years. The results are presented in Table 7 (the silver standard) and in Table 8 (the gold standard).

While reducing the Meteor tables during the evaluation is still beneficial, this is not entirely valid about the prior paraphrasing of reference sentences. The baseline correlation of Meteor is rather high and paraphrasing sometimes helps and sometimes harms the performance of the metric. Nevertheless, the differences in correlation between the original references and the new ones are very small (0.012 at most).

In contrast to WMT12 and WMT13, the first phase of paraphrasing before applying Depfix causes a drop in correlation. On the other hand, applying Depfix is again always beneficial.

With both standards, the best result is achieved on the original reference with the **No paraphr.** and the **WordNet** setting. Parmesan outperforms Meteor by a marginal difference (0.005) on the silver standard, whereas using the gold standard, Meteor is better by exactly the same margin. However, the correlation of the two standards is 0.997.

There is a distinctive difference between the data from previous years and this one. In the WMT14, the English source data for translating to Czech are sentences originally English or professionally translated from Czech to English. In the previous years, on the other hand, the source data were equally composed from all competing languages, i.e., only fifth/sixth of data is originally English.

One more language involved in the translation seems as a possible ground for the beneficial effect of prior paraphrasing of reference sentences. Therefore, we experiment with limiting the WMT12 and WMT13 data to only sentences that are originally Czech or English. However, Parmesan on this limited translations again significantly outperforms Meteor and the results (see Table 9) follow similar patterns as on the whole data sets.

8 Conclusion and Future Work

We have demonstrated a negative effect of noise in the Czech Meteor Paraphrase tables to the performance of Meteor. We have shown that large-scale reduction of the paraphrase tables can be very beneficial for targeted paraphrasing of reference sentences. The Meteor scores computed without the Czech Meteor Paraphrase tables on these new reference sentences correlates significantly better with the human judgment than original Meteor on the WMT12 and WMT13 data. However, the WMT14 data has not confirmed

WMT14

reference	Basic	One-word	Same POS	Same Lemma	Diff. Lemma	No paraphr.	WordNet
Original	0.963	0.967	0.965	0.968	0.970	**0.973**	0.973
Before Depfix	0.957	0.958	0.959	0.959	0.965	0.965	0.965
New	0.968	0.965	0.969	0.969	0.968	0.968	0.968

Table 7: Pearson's correlation of Meteor and the silver standard.

WMT14

reference	Basic	One-word	Same POS	Same Lemma	Diff. Lemma	No paraphr.	WordNet
Original	0.967	0.968	0.969	0.972	0.972	**0.974**	**0.974**
Before Depfix	0.958	0.959	0.959	0.960	0.963	0.963	0.963
New	0.966	0.966	0.966	0.967	0.962	0.962	0.962

Table 8: Pearson's correlation of Meteor and the gold standard – *TrueSkill* (Bojar et al., 2014). Note that as opposed to official WMT14 results, the version 1.4 of Meteor is still used in this table.

WMT12

reference	Basic	One-word	Same POS	Same Lemma	Diff. Lemma	No paraphr.	WordNet
Original	0.781	0.779	0.782	0.772	0.807	0.798	0.801
Before Depfix	0.872	0.872	0.874	0.874	0.898	0.899	0.899
New	0.897	0.897	0.897	0.897	**0.923**	**0.923**	**0.923**

WMT13

reference	Basic	One-word	Same POS	Same Lemma	Diff. Lemma	No paraphr.	WordNet
Original	0.805	0.810	0.813	0.813	0.842	0.840	0.844
Before Depfix	0.843	0.846	0.849	0.848	0.879	0.877	0.877
New	0.874	0.877	0.878	0.877	0.877	**0.902**	**0.902**

Table 9: Pearson's correlation of Meteor and the silver standard on sentences originally Czech or English only. In this case, the interpretation of human judgment was computed only on those sentences as well.

this result and the improvement was very small. Furthermore, Parmesan performs even worse than Meteor on the gold standard.

In the future, we plan to thoroughly examine the reason for the different performance on WMT14 data. We also intend to make more sophisticated paraphrases including word order changes and other transformation that cannot be expressed by simple substitution of two words. We are also considering extending Parmesan to more languages.

Acknowledgment

I would like to thank Ondřej Bojar for his helpful suggestions. This research was partially supported by the grants SVV project number 260 104 and FP7-ICT-2011-7-288487 (MosesCore). This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petra Barančíková, Rudolf Rosa, and Aleš Tamchyna. 2014. Improving Evaluation of English-Czech MT through Paraphrasing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavík, Iceland. European Language Resources Association.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar F. Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 1–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tam-

chyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proc. of LREC*, pages 3921–3928. ELRA.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. In *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 455–462, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiao-Li Meng, Robert Rosenthal, and Donald B Rubin. 1992. Comparing correlated correlation coefficients. *Psychological bulletin*, 111(1):172.

George A. Miller. 1995. WordNet: A Lexical Database for English. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Karel Pala and Pavel Smrž. 2004. Building Czech WordNet. *In Romanian Journal of Information Science and Technology*, 7:79–88.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 362–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha.

A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU

Boxing Chen and Colin Cherry
National Research Council Canada
first.last@nrc-cnrc.gc.ca

Abstract

BLEU is the *de facto* standard machine translation (MT) evaluation metric. However, because BLEU computes a geometric mean of n-gram precisions, it often correlates poorly with human judgment on the sentence-level. Therefore, several smoothing techniques have been proposed. This paper systematically compares 7 smoothing techniques for sentence-level BLEU. Three of them are first proposed in this paper, and they correlate better with human judgments on the sentence-level than other smoothing techniques. Moreover, we also compare the performance of using the 7 smoothing techniques in statistical machine translation tuning.

1 Introduction

Since its invention, BLEU (Papineni et al., 2002) has been the most widely used metric for both machine translation (MT) evaluation and tuning. Many other metrics correlate better with human judgments of translation quality than BLEU, as shown in recent WMT Evaluation Task reports (Callison-Burch et al., 2011; Callison-Burch et al., 2012). However, BLEU remains the *de facto* standard evaluation and tuning metric. This is probably due to the following facts:

1. BLEU is language independent (except for word segmentation decisions).
2. BLEU can be computed quickly. This is important when choosing a tuning metric.
3. BLEU seems to be the best tuning metric from a quality point of view - i.e., models trained using BLEU obtain the highest scores from humans and even from other metrics (Cer et al., 2010).

One of the main criticisms of BLEU is that it has a poor correlation with human judgments on the sentence-level. Because it computes a geometric mean of n-gram precisions, if a higher order n-gram precision (eg. $n = 4$) of a sentence is 0, then the BLEU score of the entire sentence is 0, no matter how many 1-grams or 2-grams are matched. Therefore, several smoothing techniques for sentence-level BLEU have been proposed (Lin and Och, 2004; Gao and He, 2013).

In this paper, we systematically compare 7 smoothing techniques for sentence-level BLEU. Three of them are first proposed in this paper, and they correlate better with human judgments on the sentence-level than other smoothing techniques on the WMT metrics task. Moreover, we compare the performance of using the 7 smoothing techniques in statistical machine translation tuning on NIST Chinese-to-English and Arabic-to-English tasks. We show that when tuning optimizes the expected sum of these sentence-level metrics (as advocated by Cherry and Foster (2012) and Gao and He (2013) among others), all of these metrics perform similarly in terms of their ability to produce strong BLEU scores on a held-out test set.

2 BLEU and smoothing

2.1 BLEU

Suppose we have a translation T and its reference R, BLEU is computed with precision $P(N, T, R)$ and brevity penalty BP(T,R):

$$\text{BLEU}(N, T, R) = P(N, T, R) \times \text{BP}(T, R) \quad (1)$$

where $P(N, T, R)$ is the geometric mean of n-gram precisions:

$$P(N, T, R) = \left(\prod_{n=1}^{N} p_n \right)^{\frac{1}{N}} \quad (2)$$

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 362–367,
Baltimore, Maryland USA, June 26–27, 2014.