The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 135–147

Evaluation of Machine Translation Metrics for Czech as the Target Language

Kamil Kos, Ondřej Bojar

Abstract

In the present work we study semi-automatic evaluation techniques of machine translation (MT) systems. These techniques are based on a comparison of the MT system's output to human translations of the same text. Various metrics were proposed in the recent years, ranging from metrics using only a unigram comparison to metrics that try to take advantage of additional syntactic or semantic information. The main goal of this article is to compare these metrics with respect to their correlation with human judgments for Czech as the target language and to propose the best ones that can be used for an evaluation of MT systems translating into Czech language.

1. Introduction

In recent years a lot of research has been devoted to the field of MT evaluation. Since 2002, almost every year new MT metrics emerged that tried to establish themselves as the MT evaluation standard.

So far, the BLEU metric is considered as the golden standard in various competitions and workshops. However, some researchers have noted that BLEU is not very reliable in scoring translations on the sentence level. This can be a significant problem because MT systems usually translate source text sentence by sentence. Moreover, it is easier to collect human judgments on the sentence level because people can judge the quality of translations on the sentence level more easily than for the whole text.

In this article we examine MT metrics with respect to their correlation with human judgments on the level of the sentence and the translation system as a whole. We restrict our experiments only on Czech as target language because results for English are already available in Callison-Burch et al. (2008) and Callison-Burch et al. (2007). Because Czech belongs to a typologically different group of languages, namely the

Please cite this article as: Kamil Kos, Ondřej Bojar, Evaluation of Machine Translation Metrics for Czech as the Target Language. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 135–147.

Slavic ones with rich inflection, there can be some differences in the correlation. Some of the metrics can be more suitable for English and some of them more suitable for Czech, e.g. because of the fixed word order in English and relatively free word order in Czech.

2. Metrics

We compared the most common metrics that are used in MT systems evaluation. We used our own implementation of the metrics to compute the ratings. This was especially necessary for metrics that take advantage of syntactic or semantic information because original evaluation tools are available mostly only for English or other widespread languages like French or Spanish.

The following metrics were evaluated:

- **F-measure** is defined as the harmonic mean of *precision (p)* and *recall (r)*: $\frac{p+r}{2*p*r}$ where precision is the number of words that co-occur in the candidate and the reference sentence divided by the size of the candidate sentence, and recall is the number of words that co-occur in the candidate and the reference sentence divided by the size of the reference sentence.
- **BLEU** (Papineni et al., 2002) is based on the geometric mean of n-gram precision ($n = 1 \ldots 4$). Candidate translations that are shorter than human references are penalized by the brevity penalty which is a single value over the whole test set.
- **NIST** (Doddington, 2002) also uses n-gram precision ($n = 1 \ldots 5$), differing from BLEU in that an arithmetic mean is used, weights are used to emphasize informative word sequences and the formula for brevity penalty is different.
- **WER** (Su and Wu, 1992) is defined as the minimum number of edit operations required to transform one sentence into another normalized by the length of the reference translation

$$WER(s_i, r_i) = \frac{min\,(I(s_i, r_i) + D(s_i, r_i) + S(s_i, r_i))}{|r_i|}$$

  where $I(s_i, r_i)$, $D(s_i, r_i)$ and $S(s_i, r_i)$ are the number of insertions, deletions and substitutions, respectively, and $|r_i|$ is the length of the reference. The numerator of the equation above is also known as the Levenshtein distance.
- **TER** (Snover et al., 2006) is also based on the number of operations needed to transform the candidate sentence into the reference sentence. However, it allows one additional operation: the *block shift*. Hence, possible operations include insertion, deletion, and substitution of single words as well as shifts of word sequences.
- **PER** (Tillmann et al., 1997) is similar to WER except that word order is not taken into account. Both sentences are treated as bags of words and the set difference is judged.

- **GTM** (Turian et al., 2003) is inspired by the plain F-measure trying to eliminate (one of) its major drawbacks. Since F-measure is based only on unigram matching, two sentences containing the same words always get the same F-measure rating regardless of the correct order of the words in the sentence. GTM rewards contiguous sequences of correctly translated words. The reward is controlled by parameter e. For $e = 1$ the GTM score is the same as the plain F-measure. For $0 < e < 1$ contiguous sequences of words are rewarded and for $e > 1$ they are penalized.
- **Meteor** (Banerjee and Lavie, 2005) incrementally constructs an alignment between the candidate and the reference sentence using several modules that define which words can be matched. The modules are *exact*, *porter stem* and *WordNet (WN) synonymy*. *Exact* module matches two words if they have the same surface representation (e.g. *dog* matches *dog* but not *dogs*). *Porter stem* module matches two words if they have the same stem according to Porter stemmer (Porter, 2001) (e.g. *dogs* matches *dog*) and *WN synonymy* module matches two words if they are synonyms. Our modification of the metric replaces the *porter stem* module with *lemma* module which matches two words, if they have the same lemma. The *WN synonymy* module uses the Czech WordNet (Pala and Smrž, 2004). The alignment is then used to compute precision and recall, similarly to F-measure, only that the weight of precision is bigger than the weight of recall. Moreover, penalty is used to penalize translations with words in wrong order.
  In Lavie and Agarwal (2007), the authors optimized the parameters that are used by Meteor. We use the parameters that were obtained for English because they did not consider Czech. The new parameters put more weight on recall than before and use different coefficients in the penalty formula. We denote the original version of Meteor as *orig* and the new version without any attributes.
- **Semantic POS Overlapping** (SemPOS) metric is inspired by a set of metrics using various linguistic features on syntactic and semantic level introduced by Giménez and Márquez (2007). One of their best performing metrics was *semantic role overlapping*. Since we did not find a tool that would assign semantic roles as defined in Giménez and Márquez (2007) to words in a Czech sentence, we decided to use a slightly different metric. The *TectoMT* framework (Žabokrtský et al., 2008) can assign a semantic part of speech (semantic POS) to words. We compute overlapping for this linguistic feature as defined in Giménez and Márquez (2007). Moreover, we do not use the surface representation of the words but their `t-lemma` obtained from the *TectoMT* framework for the computation of the overlapping. As an approximation, we can say that our application of SemPOS evaluates the lexical choice of autosemantic words, taking the (semantic) part of speech into account.

Judgments per sentence	1	2	3	4	5	6	7	Total Sents.	Total Judgs.
Articles: # of sents.	119	24	8	3	5	3	3	165	267
Editorials: # of sents.	109	26	9	8	1	3	0	156	243

Table 1. Number of sentences with 1 to 7 human ratings in the test sets.

3. Test Data

The test data and human judgments were taken from the data collected at the Third Workshop on Statistical Machine Translation (Callison-Burch et al., 2007). We have chosen only systems and human judgments which had Czech as the target language. We used the human rankings of whole sentences. The judgments about syntactic constituents were not taken into account.

The output of the following systems was considered:

- BOJAR - Charles University, Bojar (Bojar and Hajič, 2008),
- TMT - Charles University, TectoMT (Žabokrtský et al., 2008),
- UEDIN - University of Edinburgh (Koehn et al., 2008),
- PCT - PC Translator (a commercial MT provider from the Czech Republic).

The test data consisted of two test sets. The first one contained a total of 90 articles which were selected from a variety of Czech, English, French, German, Hungarian and Spanish news sites. The other test set was drawn from Czech-English news editorials. The Articles test set contained 2050 sentences and the Editorials test set contained 2028 sentences. The reference translations contained only one human translation for each sentence.

The human judgments contained 243 system scores of 156 unique sentences for the Editorials test set and 267 system scores of 165 unique sentences for the Articles test set with up to 7 judgments of a single sentence. Table 1 gives the details of judgment distribution. The human judgments contained scores of the translation quality on the scale 1 to 5, one being the best. It was possible that several translations obtained the same score. The scores for the translations were only on the sentence level. We considered human scores of the same sentence as independent of each other and included all of them in the ratings.

4. Correlation with Human Judgments

To measure the correlation of the metric ratings with the human judgments we used the Pearson correlation coefficient on ranks. This coefficient captures the extent to which two different rankings correlate with each other. We used the following equation:

$$\rho = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}\sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}}$$

Human score	Metric score	Human rank	Metric rank
1	0.62	1.5	1
3	0.54	3	3
1	0.54	1.5	3
5	0.54	4	3

Table 2. Conversion of scores to rankings.

In the formula, n denotes the number of evaluated systems and x_i, y_i are the positions of the i^{th} system in the human and metric rank. The possible values of ρ range between 1 (all systems are ranked in the same order) and -1 (systems are ranked in the reverse order). Thus, an evaluation metric with a higher value of ρ reflects the human judgments better than a metric with a lower ρ.

4.1. Sentence-Level Correlation

To measure the sentence-level correlation we transformed the human scores to ranks for each sentence. If several systems obtained the same score, we used the average position for each of them. In the case that all systems had the same score, we did not use the human judgment. For automatic metrics, we computed the metric scores on the sentence level and converted the scores to rankings in the same manner as for human judgments. Table 2 illustrates how we created the rankings.

4.2. System-Level Correlation

Because no human judgments were available on the system level we had to synthesize them from sentence level judgments. We used the same method as in Callison-Burch et al. (2007) in order to make the results comparable. We created the system rankings based on the

- *percent of cases in which the sentences (produced by the system) were judged to be better than or equal to the translations of any other system.*

Since we had only two test sets to measure the correlation coefficients on the system level, we used bootstrapping to estimate their variance. On the system level, we obtained no ties in rankings. Then, the Pearson correlation coefficient is equivalent to the Spearman's rank correlation coefficient defined as:

$$\rho_{sp} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks for $system_i$ and n is the number of systems.

Metric	Articles	Editorials	Average
NIST	0.22±0.60 (7)	0.26±0.62 (1)	0.24
F-measure/GTM(e=1)	0.24±0.58 (1)	0.23±0.63 (4)	0.23
GTM(e=0.5)	0.24±0.58 (2)	0.23±0.63 (6)	0.23
GTM(e=2)	0.24±0.58 (3)	0.22±0.63 (10)	0.23
Meteor	0.23±0.57 (4)	0.24±0.62 (2)	0.23
GTM(e=0.1)	0.23±0.58 (5)	0.23±0.63 (5)	0.23
Meteor(orig)	0.23±0.57 (6)	0.23±0.62 (7)	0.23
PER	0.22±0.60 (8)	0.24±0.63 (3)	0.23
TER	0.21±0.60 (9)	0.23±0.62 (8)	0.22
WER	0.21±0.60 (10)	0.23±0.62 (9)	0.22
SemPOS	0.21±0.57 (11)	0.19±0.61 (11)	0.20
BLEU	0.03±0.63 (12)	0.02±0.62 (12)	0.03

Numbers in brackets indicate the relative position of the metric.

Table 3. Average sentence-level correlations for the metrics including standard deviation.

5. Results and Discussion

In the present section, we discuss various aspects of the estimated correlations to human judgments. For complete listing of results, please see Tables 7 and 8 at the end of our article.

5.1. BLEU Not Suitable for Sentence-Level Evaluation

The results of the sentence-level correlation are given in Table 3. They indicate that the correlation of the automatic metrics with human judgments is not very high (around 0.2). Perhaps more importantly, the huge variance of the correlation discards any differences between the metrics. In fact, all results lie within the error bars of the best performing metrics (NIST for the Editorials dataset and F-measure/GTM(e=1) for the Articles dataset).

The only outstanding result is the extremely low correlation for BLEU. The BLEU metric cannot predict the human judgments on the sentence level at all which makes it unsuitable for evaluation of the quality of separate sentences.

5.2. Sentence-Level Correlation Difficult for Humans

The low coefficients observed in Table 3 are, however, influenced by the quality of human judgments. The inter-human correlation coefficients are given in Table 4. They suggest that it is difficult even for human annotators to agree which sentence

	Articles	Editorials
Judgment pairs	224	156
ρ	0.56±0.48	0.56±0.50

Table 4. Number of human judgment pairs of the same sentence and the average inter-human correlations with standard deviation.

translations are good. For an illustration of two sentences see Figures 1 and 2 at the end of the paper.

The inter-human correlation coefficients were computed as follows: we took the human scores for sentences for which there were given at least two human judgments and computed the Pearson's correlation coefficient for them. If there were more than two ratings of the same sentence, we considered all possible combinations. For the Editorials test set, we obtained 156 pairs of human judgments and for the Articles test set 224 pairs.

5.3. SemPOS Best for System-Level Comparison

Table 5 presents average Pearson correlation coefficients for both test sets on the system level. We used bootstrapping to estimate the confidence intervals. We can see that the Semantic POS Overlapping metric clearly has the highest correlation, followed by the Meteor metric. The next metrics are GTM(e=0.5) and BLEU. Metrics with the lowest correlation were the distance metrics PER, WER and TER.

It is interesting that NIST, the best metric on the sentence level, finished in the second half of the chart on the system level. On the contrary, BLEU can evaluate the quality of translation much better on the system level than on the sentence level, even if it is only slightly better than the average metrics on the system level.

Note that the Semantic POS Overlapping extensively takes advantage of the automatic annotation tools. The MT output must be preprocessed first to obtain the semantic POS and `t-lemma` for the words of the translation. Hence, the performance of Semantic POS Overlapping metric can be influenced by the quality of the annotation tools.

5.4. Effects of Lemmatization

Table 6 illustrates the effects of lemmatizing both the reference and the hypothesis of the system for selected metrics. By lemmatizing, we deliberately ignore differences in word forms. The systems are therefore not judged on the basis of morphological coherence of the output.

The column "lemma" shows correlations for texts lemmatized while preserving the number of tokens. The column "t-lemma" shows correlations for linearized tec-

Metric	Articles	Editorials	Average
SemPOS	**0.81±0.18 (1)**	**0.75±0.23 (1)**	0.78
Meteor	0.43±0.18 (2)	**0.60±0.28 (2)**	0.52
Meteor(orig)	0.43±0.18 (3)	**0.52±0.32 (3)**	0.47
GTM(e=0.1)	0.24±0.34 (9)	*0.48±0.34 (4)*	0.36
GTM(e=0.5)	0.40±0.22 (5)	0.28±0.33 (5)	0.34
BLEU	0.40±0.23 (6)	0.25±0.33 (6)	0.33
F-measure/GTM(e=1)	0.41±0.21 (4)	0.21±0.31 (7)	0.31
GTM(e=2)	0.31±0.34 (7)	0.18±0.31 (9)	0.24
NIST	0.25±0.34 (8)	0.21±0.31 (8)	0.23
PER	0.01±0.38 (10)	0.16±0.32 (12)	0.09
TER	-0.17±0.41 (11)	0.18±0.32 (10)	0.00
WER	-0.17±0.41 (12)	0.18±0.32 (11)	0.00

Results covered in the error bounds of the best result are in bold. Results covering the best result in their error bounds are in italics. Numbers in brackets indicate the relative position of the metric.

Table 5. Average system-level correlations with standard deviations for the metrics computed from bootstrapped samples (N=10000).

togrammatical trees where the number of tokens has been reduced (auxiliary words are removed, the reflexive particle becomes part of the verb t-lemma).

The results are not very pronounced, the error bars always cover the differences. In general, lemmatization tends to improve the correlation but for some metrics and some datasets, the correlation can significantly drop.

As can be seen in Table 8 at the end of the paper, SemPOS remains the best performing metric for the system-level comparison. For the sentence-level comparison, lemmatization puts the very simple PER metric higher on the scale, see Table 7.

5.5. Comparison with English

If we compare our results with the correlation coefficients on the system level that were published in Callison-Burch et al. (2008) and Callison-Burch et al. (2007), we can see that the results for Czech and English as the target language are similar. Meteor and SemPOS (which is similar to Semantic Roles Overlapping (SR) metric from Callison-Burch et al., 2007) correlate the best with human judgments, while TER (mTER in Callison-Burch et al., 2007) has one of the lowest correlation coefficients. However, almost all metrics, except for SemPOS, show correlation coefficients of only 0.3 to 0.4 for Czech compared to 0.6 to 0.8 for English. We have documented that the distance metrics PER, WER and TER are completely unsuitable for system-level evaluation for Czech. We explain this by the morphological richness of Czech—various

Metric	Dataset	word form	lemma	t-lemma
BLEU	Articles	0.40±0.23	↘0.36±0.30	↘0.14±0.46
	Editorials	0.25±0.33	↗0.43±0.35	↗0.50±0.32
F-measure/GTM(e=1)	Articles	0.41±0.21	↗0.49±0.21	↗0.56±0.24
	Editorials	0.21±0.31	↗0.29±0.34	↗0.41±0.35
GTM(e=0.1)	Articles	0.24±0.34	↘-0.19±0.35	↘0.01±0.41
	Editorials	0.48±0.34	↘0.44±0.35	↗0.66±0.23
GTM(e=0.5)	Articles	0.40±0.22	↘0.39±0.23	↗0.47±0.26
	Editorials	0.28±0.33	↗0.48±0.33	↗0.62±0.25
GTM(e=2)	Articles	0.31±0.34	↗0.64±0.26	↗0.56±0.28
	Editorials	0.18±0.31	⇝0.18±0.32	↗0.21±0.32
NIST	Articles	0.25±0.34	↗0.50±0.32	↗0.32±0.36
	Editorials	0.21±0.31	↗0.32±0.35	↗0.33±0.35
PER	Articles	0.01±0.38	↗0.21±0.42	↘-0.09±0.35
	Editorials	0.16±0.32	↗0.20±0.33	↗0.19±0.33

Table 6. Effects of lemmatization on system-level correlation.

paraphrases with the same meaning can often differ in every word form due to e.g. a different case.

6. Future Work

More accurate results about the quality of MT metrics for Czech as the target language can be obtained if the experiments we have performed on the system level would be repeated on more data. We had only 2 test sets of 156 and 165 unique sentences. Since the synthesized system-level human judgments from sentence-level scores are possible sources of errors, it would be useful to collect human judgments for whole test sets. However, people are better at scoring shorter fragments of text and they even have problems when evaluating longer sentences. We suggest to use a task-based evaluation instead of scoring the whole text. Human judgments can be collected on the basis of a set of questions after the translations are read by the annotators, like in comprehension tests in foreign language exams.

Other metrics that emerged recently can be implemented and evaluated. This concerns especially metrics that were published in Giménez and Márquez (2007). Several of them show high correlation with human judgments for English. The TectoMT framework can provide most of the required features to compute these metrics for Czech sentences.

		Human scores	
Src	Berlusconi's lack of experience in politics doomed his first government to collapse after only six months.		
Ref	Berlusconiho nedostatečné zkušenosti v politice odsoudily jeho první vládu po pouhých šesti měsících k pádu.		
BOJAR	Berlusconiho nedostatek zkušeností v politice odsoudila jeho první vláda, Berlusconi's lack$_{nom,acc}$ experience$^{pl}_{gen}$ in politics doomed his first gov.$_{nom}$, která padla po pouhých šest měsíců. which collapsed after only six$_{nom}$ months.	**3**	3
TMT	Nedostatek Berlusconiho zkušenosti z politik odsoudil svou první vládu Lack of Berlusconi's experience$^{sg}_{gen}$ from politics doomed his first gov.$_{acc}$ zhroutit po pouze šesti měsících. to collapse after only six months.	4	**2**
UEDIN	Berlusconi's nedostatku zkušeností v politice odsouzena k zániku jeho první Berlusconi's lack$_{dat}$ experience$^{pl}_{gen}$ in politics doomed to downfall his first vláda, která padla po pouhých šest měsíců. gov.$_{nom}$, which collapsed after only six$_{nom}$ months.	4	4
PCT	Berlusconi- ovo nedostatek zkušenosti s politikami odsoudily jeho první Berlusconi- 's lack experience$^{sg}_{gen}$ with politics$_{pl}$ doomed$_{pl}$ his first vládu k tomu, aby zhroutila se po jen šesti měsíců. gov. to that, so that collapsed *refl* after only six months$_{gen}$.	**3**	3

Figure 1. Example sentence 1 with human scores

		Human scores				
Src	The former police chief has been cooperating fully with the prosecutors investigating the case, Morvai added.					
Ref	Attila Morvai se zmínil taktéž o tom, že bývalý policejní kapitán od začátku spolupracoval se státními zástupci vykonávajícími vyšetřování.					
BOJAR	Bývalý policejní šéf byl plné spolupráci s prokurátory Morvai Former police chief was full cooperation with prosecutors Morvai vyšetřování případu, dodal. investigation case$_{gen}$, added.	4	4	4	4	**3**
TMT	Že se bývalý policejní šéf spolupracoval plně žalobci That *refl* former police chief cooperated fully prosecutors vyšetřováním případu, Morvai přidal. investigation$_{inst}$ case$_{gen}$, Morvai added.	3	3	3	**2**	4
UEDIN	Bývalý náčelník policie bylo plné spolupráci s prokurátory Former chief police was full cooperation with prosecutors vyšetřování případu, morvai přidán. investigation case$_{gen}$, morvai added$_{pass}$.	2	3	4	4	4
PCT	Bývalý policejní šéf spolupracoval plně se žalobce Former police chief cooperated fully with prosecutor$_{nom}$ vyšetřování případ, Morvai přidal. investigation case$_{nom}$, Morvai added.	**1**	**2**	**2**	**2**	4

Figure 2. Example sentence 2 with human scores

Preprocessing	Metric	Articles	Editorials	Average
lemma	PER	**0.24±0.57 (1)**	**0.28±0.61 (2)**	0.26
t-lemma	PER	**0.21±0.56 (17)**	**0.30±0.59 (1)**	0.26
lemma	F-measure/GTM(e=1)	**0.24±0.58 (2)**	**0.24±0.60 (14)**	0.24
t-lemma	NIST	**0.24±0.56 (3)**	**0.24±0.58 (15)**	0.24
–	NIST	**0.22±0.60 (11)**	**0.26±0.62 (3)**	0.24
t-lemma	F-measure/GTM(e=1)	**0.22±0.57 (12)**	**0.26±0.59 (6)**	0.24
t-lemma	GTM(e=0.1)	**0.22±0.57 (13)**	**0.26±0.59 (7)**	0.24
t-lemma	GTM(e=0.5)	**0.22±0.57 (14)**	**0.26±0.59 (8)**	0.24
–	F-measure/GTM(e=1)	**0.24±0.58 (4)**	**0.23±0.63 (16)**	0.23
–	GTM(e=0.5)	**0.24±0.58 (5)**	**0.23±0.63 (18)**	0.23
–	GTM(e=2)	**0.24±0.58 (6)**	**0.22±0.63 (24)**	0.23
–	Meteor	**0.23±0.57 (7)**	**0.24±0.62 (12)**	0.23
–	GTM(e=0.1)	**0.23±0.58 (8)**	**0.23±0.63 (17)**	0.23
–	Meteor(orig)	**0.23±0.57 (9)**	**0.23±0.62 (19)**	0.23
lemma	GTM(e=2)	**0.23±0.59 (10)**	**0.23±0.62 (23)**	0.23
–	PER	**0.22±0.60 (15)**	**0.24±0.63 (13)**	0.23
lemma	GTM(e=0.5)	**0.22±0.59 (16)**	**0.23±0.60 (22)**	0.23
t-lemma	GTM(e=2)	**0.21±0.57 (18)**	**0.26±0.59 (9)**	0.23
lemma	TER	**0.19±0.57 (24)**	**0.26±0.61 (4)**	0.23
lemma	WER	**0.19±0.57 (25)**	**0.26±0.61 (5)**	0.23
–	TER	**0.21±0.60 (19)**	**0.23±0.62 (20)**	0.22
–	WER	**0.21±0.60 (20)**	**0.23±0.62 (21)**	0.22
lemma	GTM(e=0.1)	**0.21±0.60 (21)**	**0.22±0.59 (25)**	0.21
lemma	NIST	**0.21±0.59 (22)**	**0.22±0.61 (26)**	0.21
–	SemPOS	**0.21±0.57 (23)**	**0.19±0.61 (27)**	0.20
t-lemma	TER	**0.13±0.61 (26)**	**0.25±0.62 (10)**	0.19
t-lemma	WER	**0.13±0.61 (27)**	**0.25±0.62 (11)**	0.19
lemma	BLEU	**0.09±0.60 (28)**	**0.02±0.64 (30)**	0.06
t-lemma	BLEU	**0.02±0.58 (30)**	**0.06±0.63 (28)**	0.04
–	BLEU	**0.03±0.63 (29)**	**0.02±0.62 (29)**	0.03

Results covered in the error bounds of the best result in bold.

Table 7. Sentence-level correlations with human judgments.

Preprocessing	Metric	Articles	Editorials	Average
–	SemPOS	**0.81±0.18 (1)**	**0.75±0.23 (1)**	0.78
t-lemma	GTM(e=0.5)	0.47±0.26 (7)	**0.62±0.25 (3)**	0.54
–	Meteor	0.43±0.18 (8)	**0.60±0.28 (4)**	0.52
t-lemma	F-measure/GTM(e=1)	0.56±0.24 (3)	*0.41±0.35 (11)*	0.48
–	Meteor(orig)	0.43±0.18 (9)	**0.52±0.32 (5)**	0.47
lemma	GTM(e=0.5)	0.39±0.23 (13)	*0.48±0.33 (8)*	0.43
lemma	GTM(e=2)	**0.64±0.26 (2)**	0.18±0.32 (25)	0.41
lemma	NIST	*0.50±0.32 (5)*	0.32±0.35 (13)	0.41
lemma	BLEU	0.36±0.30 (14)	*0.43±0.35 (10)*	0.40
t-lemma	GTM(e=2)	*0.56±0.28 (4)*	0.21±0.32 (19)	0.39
lemma	F-measure/GTM(e=1)	0.49±0.21 (6)	0.29±0.34 (14)	0.39
–	GTM(e=0.1)	0.24±0.34 (18)	*0.48±0.34 (7)*	0.36
–	GTM(e=0.5)	0.40±0.22 (11)	0.28±0.33 (15)	0.34
t-lemma	GTM(e=0.1)	0.01±0.41 (21)	**0.66±0.23 (2)**	0.34
–	BLEU	0.40±0.23 (12)	0.25±0.33 (16)	0.33
t-lemma	NIST	0.32±0.36 (15)	0.33±0.35 (12)	0.33
t-lemma	BLEU	0.14±0.46 (20)	*0.50±0.32 (6)*	0.32
–	F-measure/GTM(e=1)	0.41±0.21 (10)	0.21±0.31 (17)	0.31
–	GTM(e=2)	0.31±0.34 (16)	0.18±0.31 (22)	0.24
–	NIST	0.25±0.34 (17)	0.21±0.31 (18)	0.23
lemma	PER	0.21±0.42 (19)	0.20±0.33 (20)	0.21
lemma	GTM(e=0.1)	-0.19±0.35 (30)	*0.44±0.35 (9)*	0.12
–	PER	0.01±0.38 (22)	0.16±0.32 (28)	0.09
lemma	TER	-0.01±0.36 (23)	0.18±0.32 (26)	0.08
lemma	WER	-0.01±0.36 (24)	0.17±0.32 (27)	0.08
t-lemma	PER	-0.09±0.35 (25)	0.19±0.33 (21)	0.05
–	TER	-0.17±0.41 (28)	0.18±0.32 (23)	0.00
–	WER	-0.17±0.41 (29)	0.18±0.32 (24)	0.00
t-lemma	TER	-0.16±0.32 (26)	0.12±0.33 (29)	-0.02
t-lemma	WER	-0.16±0.32 (27)	0.12±0.33 (30)	-0.02

Results covered in the error bounds of the best result in bold.
Results covering the best result in their error bounds in italics.

Table 8. System-level correlations with human judgments.

7. Conclusion

This work has examined the most common MT system evaluation metrics that are currently used. The experiments have demonstrated that the most suitable metrics for evaluation of MT systems on the system level with Czech as the target language are Semantic POS Overlapping and Meteor, followed by GTM, BLEU and NIST. These results are consistent with data that were published for systems with English as the target language even though the correlation coefficients with human judgments are lower for Czech.

The evaluation of MT quality on the sentence level proved to be unsuitable because of a relatively low correlation with human judgments for all considered metrics. Due to the variance of the correlations, none of the metrics was identified as the best one. We only found out that BLEU does not correlate with human judgments on the sentence level. However, the results were influenced by the quality of human judgments which had only a moderate inter-human correlation.

8. Acknowledgment

The work on this project was supported by the grant FP6-IST-5-034291-STP (EuroMatrix), and the grants MSM0021620838 and ME838.

Bibliography

Banerjee, S. and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, pages 65–72, Ann Arbor, Michigan, June 2005.

Bojar, Ondřej and Jan Hajič. Phrase-based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, 2007. Association for Computational Linguistics.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, 2008. Association for Computational Linguistics.

Doddington, George. 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, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Giménez, Jesús and Lluís Márquez. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, June 2007. Association for Computational Linguistics.

Koehn, Philipp, Abhishek Arun, and Hieu Hoang. Towards Better Machine Translation Quality for the German-English Language Pairs. In *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Lavie, A. and A. Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, June 2007. Association for Computational Linguistics.

Pala, Karel and Pavel Smrž. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 2004(7):79–88, 2004. URL `http://www.fit.vutbr.cz/research/view_pub.php?id=7682`.

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

Porter, Martin. The Porter Stemming Algorithm, 2001. URL `http://www.tartarus.org/martin/PorterStemmer/index.html`. Last visited on July 16, 2008.

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

Su, K. and J. Wu. A New Quantitative Quality Measure for Machine Translation Systems. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 433–439, Nantes, France, July 1992.

Tillmann, Christoph, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accelerated DP Based Search for Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece, September 1997.

Turian, Joseph P., Luke Shen, and I. Dan Melamed. Evaluation of Machine Translation and its Evaluation. In *Machine Translation Summit IX*, pages 386–393. International Association for Machine Translation, September 2003.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, June 2008. Association for Computational Linguistics.