Tackling Sparse Data Issue in Machine Translation Evaluation *

Ondřej Bojar, Kamil Kos, and David Mareček
Charles University in Prague, Institute of Formal and Applied Linguistics
{bojar,marecek}@ufal.mff.cuni.cz, kamilkos@email.cz

Abstract

We illustrate and explain problems of n-grams-based machine translation (MT) metrics (e.g. BLEU) when applied to morphologically rich languages such as Czech. A novel metric SemPOS based on the deep-syntactic representation of the sentence tackles the issue and retains the performance for translation to English as well.

1 Introduction

Automatic metrics of machine translation (MT) quality are vital for research progress at a fast pace. Many automatic metrics of MT quality have been proposed and evaluated in terms of correlation with human judgments while various techniques of manual judging are being examined as well, see e.g. MetricsMATR08 (Przybocki et al., 2008)[1], WMT08 and WMT09 (Callison-Burch et al., 2008; Callison-Burch et al., 2009)[2].

The contribution of this paper is twofold. Section 2 illustrates and explains severe problems of a widely used BLEU metric (Papineni et al., 2002) when applied to Czech as a representative of languages with rich morphology. We see this as an instance of the sparse data problem well known for MT itself: too much detail in the formal representation leading to low coverage of e.g. a translation dictionary. In MT evaluation, too much detail leads to the lack of comparable parts of the hypothesis and the reference.

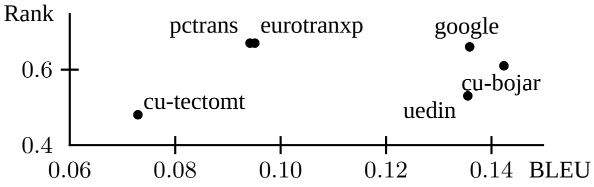

Figure 1: BLEU and human ranks of systems participating in the English-to-Czech WMT09 shared task.

Section 3 introduces and evaluates some new variations of SemPOS (Kos and Bojar, 2009), a metric based on the deep syntactic representation of the sentence performing very well for Czech as the target language. Aside from including dependency and n-gram relations in the scoring, we also apply and evaluate SemPOS for English.

2 Problems of BLEU

BLEU (Papineni et al., 2002) is an established language-independent MT metric. Its correlation to human judgments was originally deemed high (for English) but better correlating metrics (esp. for other languages) were found later, usually employing language-specific tools, see e.g. Przybocki et al. (2008) or Callison-Burch et al. (2009). The unbeaten advantage of BLEU is its simplicity.

Figure 1 illustrates a very low correlation to human judgments when translating to Czech. We plot the official BLEU score against the rank established as the percentage of sentences where a system ranked no worse than all its competitors (Callison-Burch et al., 2009). The systems developed at Charles University (cu-) are described in Bojar et al. (2009), uedin is a vanilla configuration of Moses (Koehn et al., 2007) and the remaining ones are commercial MT systems.

In a manual analysis, we identified the reasons for the low correlation: BLEU is overly sensitive to *sequences* and *forms* in the hypothesis matching

* This work has been supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), FP7-ICT-2009-4-247762 (Faust), GA201/09/H057, GAUK 1163/2010, and MSM 0021620838. We are grateful to the anonymous reviewers for further research suggestions.

[1] http://nist.gov/speech/tests/metricsmatr/2008/results/

[2] http://www.statmt.org/wmt08 and wmt09

Proceedings of the ACL 2010 Conference Short Papers, pages 86–91,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

Confirmed	Error Flags	1-grams	2-grams	3-grams	4-grams
Yes	Yes	6.34%	1.58%	0.55%	0.29%
Yes	No	36.93%	13.68%	5.87%	2.69%
No	Yes	22.33%	41.83%	54.64%	63.88%
No	No	**34.40%**	**42.91%**	**38.94%**	**33.14%**
Total n-grams		35,531	33,891	32,251	30,611

Table 1: n-grams confirmed by the reference and containing error flags.

the reference translation. This focus goes directly against the properties of Czech: relatively free word order allows many permutations of words and rich morphology renders many valid word forms not confirmed by the reference.[3] These problems are to some extent mitigated if several reference translations are available, but this is often not the case.

Figure 2 illustrates the problem of "sparse data" in the reference. Due to the lexical and morphological variance of Czech, only a single word in each hypothesis matches a word in the reference. In the case of pctrans, the match is even a false positive, "do" (to) is a preposition that should be used for the "minus" phrase and not for the "end of the day" phrase. In terms of BLEU, both hypotheses are equally poor but 90% of their tokens were not evaluated.

Table 1 estimates the overall magnitude of this issue: For 1-grams to 4-grams in 1640 instances (different MT outputs and different annotators) of 200 sentences with manually flagged errors[4], we count how often the n-gram is confirmed by the reference and how often it contains an error flag. The suspicious cases are n-grams confirmed by the reference but still containing a flag (false positives) and n-grams not confirmed despite containing no error flag (false negatives).

Fortunately, there are relatively few false positives in n-gram based metrics: 6.3% of unigrams and far fewer higher n-grams.

The issue of false negatives is more serious and confirms the problem of sparse data if only one reference is available. 30 to 40% of n-grams do not contain any error and yet they are not confirmed by the reference. This amounts to 34% of running unigrams, giving enough space to differ in human judgments and still remain unscored.

Figure 3 documents the issue across languages: the lower the BLEU score itself (i.e. fewer confirmed n-grams), the lower the correlation to human judgments regardless of the target language (WMT09 shared task, 2025 sentences per language).

Figure 4 illustrates the overestimation of scores caused by too much attention to sequences of tokens. A phrase-based system like Moses (cu-bojar) can sometimes produce a long sequence of tokens exactly as required by the reference, leading to a high BLEU score. The framed words in the illustration are not confirmed by the reference, but the actual error in these words is very severe for comprehension: nouns were used twice instead of finite verbs, and a misleading translation of a preposition was chosen. The output by pctrans preserves the meaning much better despite not scoring in either of the finite verbs and producing far shorter confirmed sequences.

3 Extensions of SemPOS

SemPOS (Kos and Bojar, 2009) is inspired by metrics based on overlapping of linguistic features in the reference and in the translation (Giménez and Márquez, 2007). It operates on so-called "tectogrammatical" (deep syntactic) representation of the sentence (Sgall et al., 1986; Hajič et al., 2006), formally a dependency tree that includes only autosemantic (content-bearing) words.[5] SemPOS as defined in Kos and Bojar (2009) disregards the syntactic structure and uses the semantic part of speech of the words (noun, verb, etc.). There are 19 fine-grained parts of speech. For each semantic part of speech t, the overlapping $\mathrm{O}(t)$ is set to zero if the part of speech does not occur in the reference or the candidate set and otherwise it is computed as given in Equation 1 below.

[3] Condon et al. (2009) identify similar issues when evaluating translation to Arabic and employ rule-based normalization of MT output to improve the correlation. It is beyond the scope of this paper to describe the rather different nature of morphological richness in Czech, Arabic and also other languages, e.g. German or Finnish.

[4] The dataset with manually flagged errors is available at `http://ufal.mff.cuni.cz/euromatrixplus/`

[5] We use TectoMT (Žabokrtský and Bojar, 2008), `http://ufal.mff.cuni.cz/tectomt/`, for the linguistic pre-processing. While both our implementation of SemPOS as well as TectoMT are in principle freely available, a stable public version has yet to be released. Our plans include experiments with approximating the deep syntactic analysis with a simple tagger, which would also decrease the installation burden and computation costs, at the expense of accuracy.

SRC	Prague Stock Market falls to minus by the end of the trading day
REF	pražská burza se ke konci obchodování propadla do minusu
cu-bojar	praha stock market klesne k minus na konci obchodního dne
pctrans	praha trh cenných papírů padá minus do konce obchodního dne

Figure 2: Sparse data in BLEU evaluation: Large chunks of hypotheses are not compared at all. Only a single unigram in each hypothesis is confirmed in the reference.

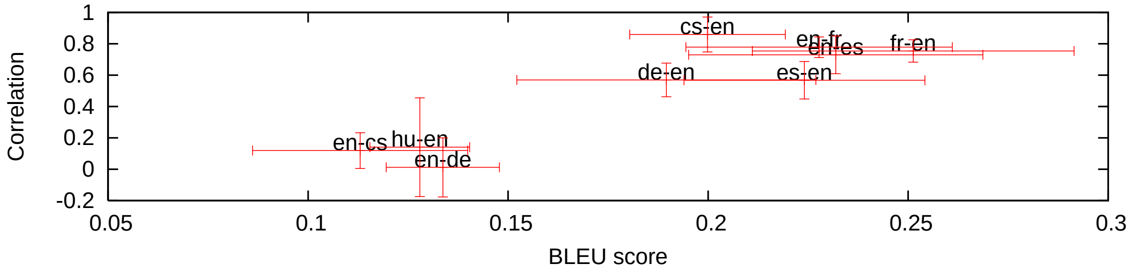

Figure 3: BLEU correlates with its correlation to human judgments. BLEU scores around 0.1 predict little about translation quality.

$$\mathrm{O}(t) = \frac{\sum_{i \in I} \sum_{w \in r_i \cap c_i} \min(\mathrm{cnt}(w,t,r_i), \mathrm{cnt}(w,t,c_i))}{\sum_{i \in I} \sum_{w \in r_i \cup c_i} \max(\mathrm{cnt}(w,t,r_i), \mathrm{cnt}(w,t,c_i))} \qquad (1)$$

The semantic part of speech is denoted t; c_i and r_i are the candidate and reference translations of sentence i, and $\mathrm{cnt}(w, t, rc)$ is the number of words w with type t in rc (the reference or the candidate). The matching is performed on the level of lemmas, i.e. no morphological information is preserved in ws. See Figure 5 for an example; the sentence is the same as in Figure 4.

The final SemPOS score is obtained by macro-averaging over all parts of speech:

$$\mathrm{SemPOS} = \frac{1}{|T|} \sum_{t \in T} \mathrm{O}(t) \qquad (2)$$

where T is the set of all possible semantic parts of speech types. (The degenerate case of blank candidate and reference has SemPOS zero.)

3.1 Variations of SemPOS

This section describes our modifications of SemPOS. All methods are evaluated in Section 3.2.

Different Classification of Autosemantic Words. SemPOS uses semantic parts of speech to classify autosemantic words. The tectogrammatical layer offers also a feature called *Functor* describing the relation of a word to its governor similarly as semantic roles do. There are 67 functor types in total.

Using *Functor* instead of *SemPOS* increases the number of word classes that independently require a high overlap. For a contrast we also completely remove the classification and use only one global class (*Void*).

Deep Syntactic Relations in SemPOS. In SemPOS, an autosemantic word of a class is confirmed if its lemma matches the reference. We utilize the dependency relations at the tectogrammatical layer to validate valence by refining the overlap and requiring also the lemma of 1) the parent (denoted “par”), or 2) all the children regardless of their order (denoted “sons”) to match.

Combining BLEU and SemPOS. One of the major drawbacks of *SemPOS* is that it completely ignores word order. This is too coarse even for languages with relatively free word order like Czech. Another issue is that it operates on lemmas and it completely disregards correct word forms. Thus, a weighted linear combination of SemPOS and BLEU (computed on the surface representation of the sentence) should compensate for this. For the purposes of the combination, we compute BLEU *only* on unigrams up to fourgrams (denoted BLEU_1, …, BLEU_4) but including the brevity penalty as usual. Here we try only a few weight settings in the linear combination but given a held-out dataset, one could optimize the weights for the best performance.

SRC	Congress yields: US government can pump 700 billion dollars into banks
REF	kongres ustoupil : vláda usa může do bank napumpovat 700 miliard dolarů
cu-bojar	kongres výnosy : vláda usa může čerpadlo 700 miliard dolarů v bankách
pctrans	kongres vynáší : us vláda může čerpat 700 miliardu dolarů do bank

Figure 4: Too much focus on sequences in BLEU: pctrans' output is better but does not score well. BLEU gave credit to cu-bojar for 1, 3, 5 and 8 fourgrams, trigrams, bigrams and unigrams, resp., but only for 0, 0, 1 and 8 n-grams produced by pctrans. Confirmed sequences of tokens are underlined and important errors (not considered by BLEU) are framed.

REF	kongres/n ustoupit/v :/n vláda/n usa/n banka/n napumpovat/v 700/n miliarda/n dolar/n
cu-bojar	kongres/n výnos/n :/n vláda/n usa/n moci/v čerpadlo/n 700/n miliarda/n dolar/n banka/n
pctrans	kongres/n vynášet/v :/n us/n vláda/n čerpat/v 700/n miliarda/n dolar/n banka/n

Figure 5: SemPOS evaluates the overlap of lemmas of autosemantic words given their semantic part of speech (n, v, . . .). Underlined words are confirmed by the reference.

SemPOS for English. The tectogrammatical layer is being adapted for English (Cinková et al., 2004; Hajič et al., 2009) and we are able to use the available tools to obtain all SemPOS features for English sentences as well.

3.2 Evaluation of SemPOS and Friends

We measured the metric performance on data used in MetricsMATR08, WMT09 and WMT08. For the evaluation of metric correlation with human judgments at the system level, we used the Pearson correlation coefficient ρ applied to ranks. In case of a tie, the systems were assigned the average position. For example if three systems achieved the same highest score (thus occupying the positions 1, 2 and 3 when sorted by score), each of them would obtain the average rank of $2 = \frac{1+2+3}{3}$. When correlating ranks (instead of exact scores) and with this handling of ties, the Pearson coefficient is equivalent to Spearman's rank correlation coefficient.

The MetricsMATR08 human judgments include preferences for pairs of MT systems saying which one of the two systems is better, while the WMT08 and WMT09 data contain system scores (for up to 5 systems) on the scale 1 to 5 for a given sentence. We assigned a human ranking to the systems based on the percent of time that their translations were judged to be better than or equal to the translations of any other system in the manual evaluation. We converted automatic metric scores to ranks.

Metrics' performance for translation to English and Czech was measured on the following testsets (the number of human judgments for a given source language in brackets):

To English: MetricsMATR08 (cn+ar: 1652), WMT08 News Articles (de: 199, fr: 251), WMT08 Europarl (es: 190, fr: 183), WMT09 (cz: 320, de: 749, es: 484, fr: 786, hu: 287)

To Czech: WMT08 News Articles (en: 267), WMT08 Commentary (en: 243), WMT09 (en: 1425)

The MetricsMATR08 testset contained 4 reference translations for each sentence whereas the remaining testsets only one reference.

Correlation coefficients for English are shown in Table 2. The best metric is Void_{par} closely followed by $\text{Void}_{\text{sons}}$. The explanation is that Void compared to SemPOS or Functor does not lose points by an erroneous assignment of the POS or the functor, and that Void_{par} profits from checking the dependency relations between autosemantic words. The combination of BLEU and SemPOS[6] outperforms both individual metrics, but in case of SemPOS only by a minimal difference. Additionally, we confirm that 4-grams alone have little discriminative power both when used as a metric of their own (BLEU_4) as well as in a linear combination with SemPOS.

The best metric for Czech (see Table 3) is a linear combination of SemPOS and 4-gram BLEU closely followed by other SemPOS and BLEU_n combinations. We assume this is because BLEU_4 can capture correctly translated fixed phrases, which is positively reflected in human judgments. Including BLEU_1 in the combination favors translations with word forms as expected by the refer-

[6] For each $n \in \{1, 2, 3, 4\}$, we show only the best weight setting for SemPOS and BLEU_n.

Metric	Avg	Best	Worst
Void$_\text{par}$	0.75	0.89	0.60
Void$_\text{sons}$	0.75	0.90	0.54
Void	0.72	0.91	0.59
Functor$_\text{sons}$	0.72	1.00	0.43
GTM	0.71	0.90	0.54
4·SemPOS+1·BLEU$_2$	0.70	0.93	0.43
SemPOS$_\text{par}$	0.70	0.93	0.30
1·SemPOS+4·BLEU$_3$	0.70	0.91	0.26
4·SemPOS+1·BLEU$_1$	0.69	0.93	0.43
NIST	0.69	0.90	0.53
SemPOS$_\text{sons}$	0.69	0.94	0.40
SemPOS	0.69	0.95	0.30
2·SemPOS+1·BLEU$_4$	0.68	0.91	0.09
BLEU$_1$	0.68	0.87	0.43
BLEU$_2$	0.68	0.90	0.26
BLEU$_3$	0.66	0.90	0.14
BLEU	0.66	0.91	0.20
TER	0.63	0.87	0.29
PER	0.63	0.88	0.32
BLEU$_4$	0.61	0.90	-0.31
Functor$_\text{par}$	0.57	0.83	-0.03
Functor	0.55	0.82	-0.09

Table 2: Average, best and worst system-level correlation coefficients for translation to English from various source languages evaluated on 10 different testsets.

Metric	Avg	Best	Worst
3·SemPOS+1·BLEU$_4$	0.55	0.83	0.14
2·SemPOS+1·BLEU$_2$	0.55	0.83	0.14
2·SemPOS+1·BLEU$_1$	0.53	0.83	0.09
4·SemPOS+1·BLEU$_3$	0.53	0.83	0.09
SemPOS	0.53	0.83	0.09
BLEU$_2$	0.43	0.83	0.09
SemPOS$_\text{par}$	0.37	0.53	0.14
Functor$_\text{sons}$	0.36	0.53	0.14
GTM	0.35	0.53	0.14
BLEU$_4$	0.33	0.53	0.09
Void	0.33	0.53	0.09
NIST	0.33	0.53	0.09
Void$_\text{sons}$	0.33	0.53	0.09
BLEU	0.33	0.53	0.09
BLEU$_3$	0.33	0.53	0.09
BLEU$_1$	0.29	0.53	-0.03
SemPOS$_\text{sons}$	0.28	0.42	0.03
Functor$_\text{par}$	0.23	0.40	0.14
Functor	0.21	0.40	0.09
Void$_\text{par}$	0.16	0.53	-0.08
PER	0.12	0.53	-0.09
TER	0.07	0.53	-0.23

Table 3: System-level correlation coefficients for English-to-Czech translation evaluated on 3 different testsets.

ence, thus allowing to spot bad word forms. In all cases, the linear combination puts more weight on SemPOS. Given the negligible difference between SemPOS alone and the linear combinations, we see that word forms are not the major issue for humans interpreting the translation—most likely because the systems so far often make more important errors. This is also confirmed by the observation that using BLEU alone is rather unreliable for Czech and BLEU-1 (which judges unigrams only) is even worse. Surprisingly BLEU-2 performed better than any other n-grams for reasons that have yet to be examined. The error metrics PER and TER showed the lowest correlation with human judgments for translation to Czech.

4 Conclusion

This paper documented problems of single-reference BLEU when applied to morphologically rich languages such as Czech. BLEU suffers from a sparse data problem, unable to judge the quality of tokens not confirmed by the reference. This is confirmed for other languages as well: the lower the BLEU score the lower the correlation to human judgments.

We introduced a refinement of SemPOS, an automatic metric of MT quality based on deep-syntactic representation of the sentence tackling the sparse data issue. SemPOS was evaluated on translation to Czech and to English, scoring better than or comparable to many established metrics.

References

Ondřej Bojar, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. 2009. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

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

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece. Association for Computational Linguistics.

Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2004. Annotation of English on the tectogrammatical level. Technical Report TR-2006-35, ÚFAL/CKL, Prague, Czech Republic, December.

Sherri Condon, Gregory A. Sanders, Dan Parvaz, Alan Rubenstein, Christy Doran, John Aberdeen, and Beatrice Oshika. 2009. Normalization for Automated Metrics: English and Arabic Speech Translation. In *MT Summit XII*.

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

Jan Hajič, Silvie Cinková, Kristýna Čermáková, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jiří Semecký, Jana Šindlerová, Josef Toman, Kristýna Tomšů, Matěj Korvas, Magdaléna Rysová, Kateřina Veselovská, and Zdeněk Žabokrtský. 2009. Prague English Dependency Treebank 1.0. Institute of Formal and Applied Linguistics, Charles University in Prague, ISBN 978-80-904175-0-2, January.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Kamil Kos and Ondřej Bojar. 2009. Evaluation of Machine Translation Metrics for Czech as the Target Language. *Prague Bulletin of Mathematical Linguistics*, 92.

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

M. Przybocki, K. Peterson, and S. Bronsart. 2008. Official results of the NIST 2008 ”Metrics for MAchine TRanslation” Challenge (MetricsMATR08).

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Zdeněk Žabokrtský and Ondřej Bojar. 2008. TectoMT, Developer's Guide. Technical Report TR-2008-39, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, December.