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.

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

\[ \text{Figure 1: BLEU and human ranks of systems participating in the English-to-Czech WMT09 shared task.} \]

\[ \text{Section 3 introduces and evaluates some new variations of SemPOS (Kos and Bojar, 2009), a metric based on the deep syntactic representation of the sentence performing very well for Czech as the target language.} \]

\[ \text{Aside from including dependency and \( n \)-gram relations in the scoring, we also apply and evaluate SemPOS for English.} \]
Table 1: n-grams confirmed by the reference and containing error flags.

<table>
<thead>
<tr>
<th>Confirmed Flags</th>
<th>1-grams</th>
<th>2-grams</th>
<th>3-grams</th>
<th>4-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes Yes</td>
<td>6.34%</td>
<td>1.58%</td>
<td>0.55%</td>
<td>0.29%</td>
</tr>
<tr>
<td>Yes No</td>
<td>36.93%</td>
<td>13.68%</td>
<td>5.87%</td>
<td>2.69%</td>
</tr>
<tr>
<td>No Yes</td>
<td>22.33%</td>
<td>41.83%</td>
<td>54.64%</td>
<td>63.88%</td>
</tr>
<tr>
<td>No No</td>
<td>34.40%</td>
<td>42.91%</td>
<td>38.94%</td>
<td>33.14%</td>
</tr>
<tr>
<td>Total n-grams</td>
<td>35,531</td>
<td>33,891</td>
<td>32,251</td>
<td>30,611</td>
</tr>
</tbody>
</table>

Table 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; Hajic 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 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.
Prague Stock Market falls to minus by the end of the trading day

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.

\[
O(t) = \frac{1}{|I|} \sum_{i \in I} \sum_{w \in r_i \cap c_i} \min(\text{cnt}(w, t, r_i), \text{cnt}(w, t, c_i))
- \frac{1}{|I|} \sum_{i \in I} \sum_{w \in r_i \cup c_i} \max(\text{cnt}(w, t, r_i), \text{cnt}(w, t, c_i))
\]

The semantic part of speech is denoted \( t \); \( c_i \) and \( r_i \) are the candidate and reference translations of sentence \( i \), and \( \text{cnt}(w, t, r_c) \) is the number of words \( w \) with type \( t \) in \( r_c \) (the reference or the candidate). The matching is performed on the level of lemmas, i.e., no morphological information is preserved in \( w \). 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:

\[
\text{SemPOS} = \frac{1}{|T|} \sum_{t \in T} O(t)
\]

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 \( \text{Functor} \) describing the relation of a word to its governor similarly as semantic roles do. There are 67 functor types in total.

Using \( \text{Functor} \) 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 (\( \text{Void} \)).

**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 “\( \text{par} \)”), or 2) all the children regardless of their order (denoted “\( \text{sons} \)”) 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 \( \text{BLEU}_{1} \) only on unigrams up to fourgrams (denoted \( \text{BLEU}_{1}, \ldots, \text{BLEU}_{4} \)) 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.
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.

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, WMT08 and WMT09. 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 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 Voidpar closely followed by Voidsons. 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 Voidpar profits from checking the dependency relations between autosemantic words. The combination of BLEU and SemPOS6 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 (BLEUl) 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 BLEUn combinations. We assume this is because BLEU4 can capture correctly translated fixed phrases, which is positively reflected in human judgments. Including BLEU1 in the combination favors translations with word forms as expected by the refer-

6For each $n \in \{1, 2, 3, 4\}$, we show only the best weight setting for SemPOS and BLEUn.
Table 2: Average, best and worst system-level correlation coefficients for translation to English from various source languages evaluated on 10 different testsets.

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

The sparse data issue. SemPOS was evaluated on translation to Czech and to English, scoring better than or comparable to many established metrics.

References


