Problems of Machine Translation Evaluation

Vít Baisa

NLP Laboratory,
Faculty of Informatics, Masaryk University,
Brno, Czech Republic
xbaisa@fi.muni.cz

Abstract. In this article we deal with general aspects of machine translation evaluation. We describe several commonly used methods of the evaluation and discuss their problems and shortcomings. Then we outline a few thoughts and ideas which try to solve mentioned problems and stand behind a design of a new method of machine translation evaluation.

Key words: machine translation; evaluation

1 Introduction

The main goal of an evaluation of machine translation (MT) is to compare different MT systems. Since we have a lot of them at our disposal (Google Translate, Yahoo Babel fish, SYSTRAN Translator, PC Translator,...) we want to be able to say which one is the best. The second, and often neglected, goal is to measure quality of translation of an arbitrary sentence.

What does a good, appropriate translation mean? There isn't the only definition but it is quite natural to expect that translation t of a sentence s should preserve the meaning of s and that t is understandable enough. The former requirement is called *accuracy* (sometimes *adequacy*) of translation whereas the latter is called *intelligibility* (sometimes *fluency*).

It is tricky to make a metric for these two features. In a human evaluation, seven or five degree scales are typically used (from the worst to the best translations), several aspects of translation are rated and, in spite of being influenced by subjectivity of human evaluators, this method is considered as the best approach. Unfortunately, it costs a lot of money and it takes a lot of time to employ people and let them manually evaluate thousands of sentences. The aim of automatic evaluation methods is to eliminate these shortcomings of the human evaluation, however, at the expense of losing exactness in the sense of correlation with the human evaluation. In other words, we look for a method which will approximate the human evaluation best.

2 Paradox of an evaluation

This searching is made difficult by a significant factor. Let us call this factor *paradox of an evaluation*. It has two points of view.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 17–**??**, 2009.

The first point could be expressed in this way: an evaluator of any translation must have better knowledge about both source and resultant language than its translator has. Otherwise he (the evaluator) would never be able to detect a single mistake of the translator.

The second point results from the first point: it is impossible to make an automatic universal evaluation method which would be better than MT itself. If we had such method at our disposal it would be quite easy to use for instance a genetic algorithm. A process of translation would start with random strings (random sentences) and in each generation it would evaluate newly evolved sentences with the evaluation method and choose adepts for the next generation. At the end we would obtain a sentence which would be as good as the evaluation method. But the algorithm could then serve as a new MT which is in contradiction with presumption.

Thus, if we want to make a versatile evaluation method, we are always limited by the paradox. The only way how to avoid the paradox is to make a non-versatile evaluation method as authors of following methods do.

3 Evaluating methods – BLEU and the others

In this section we will describe several commonly used MT evaluation methods. All these methods use referential translations of source sentences from a test set. These translations are prepared manually and, in most cases, there are several different translations of a single source sentence from test set made by several different translators.

A little more formally: we have a source sentence s, a set of its referential translations $R = \{r_i\}$ and a candidate translation c translated by a MT system from s. Evaluation methods use the candidate translation c and all referential translations r_i.

3.1 BLEU, [1] and NEVA, [2]

BLEU is the oldest evaluation method and that is probably why it is considered to be a standard. BLEU tries to find out what sentences c and r_i have in common employing n-grams. Very simply said, BLEU counts matched uni-, bi-, tri- and quadrigrams between c and r_i.

Since it is supposed that a good candidate translation should be approximately as long as a referential translation, BLEU introduces penalty for brevity. The shorter c is the worse resulting score it obtains. On the contrary, c which is longer than referential translations is implicitly penalized by n-gram matching itself.

Because BLEU fails on evaluation of short sentences, authors of the next method NEVA slightly altered BLEU's formula to achieve robustness even on short sentences. In the other aspects NEVA is very similar to BLEU.

3.2 WAFT, [2] and TER, [3]

The next two methods use *edit distance* between c and R instead of n-grams.

The method WAFT defines edit distance between c and R as minimum number of deletions, substitutions and insertions of words needed to turn c into one of $r_i \in R$ (the closest one, in the sense of edit distance, is taken and evaluated). Once edit distance is computed and normalized it serves as the evaluation of c.

Another method TER uses almost the same formula as WAFT but it works with (continuous) sequences of words. Thus we can shift (but not delete and insert) two neighbouring words in one edit step whereas in WAFT we need two edit steps to do it.

3.3 METEOR, [4]

The last method stands a bit apart from the others since it uses (as the only one) synonyms in process of an evaluation. METEOR tries to map words (unigrams) from c onto words from r_i (for all i). A word w_c can be mapped onto a word w_{r_i} if $w_c = w_{r_i}$ or w_c is synonym of w_{r_i}. METEOR exploits WordNet to be able to work with synonyms.

The whole process of the evaluation is more complex but isn't so important for us. Important thing is that, thanks to synonyms, authors of METEOR achieved higher correlation with human evaluation requiring less referential translations then the others methods.

For more details on described methods and for examples see References.

4 Problems and shortcomings of described methods

4.1 Problems concerning n-gram matching

The main idea behind usage of n-grams in machine translation evaluation is that a good candidate translation is supposed to be similar to a referential (manually prepared and thus sufficiently proper) translation. It obviously holds but what about good candidate translations differing from all referential translations? The idea strongly depends on amount of referential translations. It is evident that the more referential translations we have the higher score in the evaluation we obtain.

Unigram matching corresponds with accuracy: if we find a word in c and the same word in r_i it is probably well translated word. N-gram matching corresponds with intelligibility: human translation r_i has always high intelligibility and the longer part of r_i we match in c the higher intelligibility of c can be expected.

It has also been shown in [5] that a high score (as a result of a method which uses n-grams) probably indicates a good translation but a low score is not necessarily an indication of a poor translation.

4.2 Problems concerning edit distance

Methods using edit distance don't take relevancy of an edit step (mistake) into account at all. But it is obvious that there are a lot of possible types of mistakes differing in relevancy. Especially for morphologically rich languages, a small alteration on character level (at the end of a word as for Czech language) has smaller impact on accuracy of translation than a change of a whole word despite both are considered as one undistinguishable edit step.

Simple example proves it. The sentence *Petr mít velký červený kniha*, consisting only of Czech lemmas, has relatively high accuracy and even quite hight intelligibility (depending on a source sentence, of course, in this case s = *Peter has a big red book.*). But it would require four substitutions to change it into one of possible referential translations e.g. *Petr má velkou červenou knihu*.

Another problem concerning edit distance is complexity of computing edit distance in general.

4.3 A source sentence matters

None of presented methods takes a source sentence s into account. Since the manually prepared referential translations are supposed to be semantically equivalent (or very close) to their counterpart in a source language and also well formed, we can omit s. But not in general: it is not such a mistake to translate s wrong if s is not well structured, indeed.

One could admit that, simply, there aren't such cases of badly formed source sentences but let us consider a machine translation between *minor* languages M_1 and M_2 which requires usage (especially in statistical machine translation) of a transfer language T (typically English language). A MT system translates a sentence s_1 in language M_1 to a sentence s_t in language T but it can (and it does) make mistakes. When it translates s_t to s_2 in M_2, mistakes accumulate and the total translation is worse than the two partial translations.

So that, in this case, a quality of s_t should be taken into account for more accurate evaluation of the total translation.

5 Possible treatment

5.1 A language model: *s* and *c*

Sometimes, as we have shown, a source sentence s matters. The most straightforward way to check a quality of s is to engage language models. There are many publications about language models so we outline our idea directly.

We check every uni-, bi-, tri-, ... n-grams in s. In ideal case, the whole s would be covered by a language model. In general, the more n-grams of s are covered by the language model the better quality (in the sense of intelligibility) of s should be expected.

Checking of intelligibility of a candidate translation is much more important. The main thought behind this step is that if c is a good translation of s then c should be well formed sentence.

5.2 Semantic matching

The idea of semantic matching between s and c is similar to METEOR's mapping of words between c and r_i. The distinction is that s and c differ in languages. Thus we must exploit bilingual dictionaries. WordNet can also be used thanks to its ILI (interlingua index).

This approach brings other problems into process. Let us consider this example: $s =$ *He has a new key* and $c =$ *Má nový klíč*. It is hard task to determine a proper counterpart of a word w_s from s to a word w_c from c: both w_s and w_c can have several different meanings and we must choose the proper pair: *key* vs. *klíč* (a key for locking), *klávesa* (a key on a keyboard), *tónina* (pitch of a voice). Moreover w_s can have none counterpart: *a* (an article) and several words from s can have a single counterpart in c: *He has* and *má*.

5.3 Putting it together

Question is: what is more important – intelligibility or accuracy of a translation? It isn't easy to answer it but the goal of putting accuracy (semantic matching) and intelligibility (language model checking) together is to balance both aspects and, at the same time, dealing with intelligibility of s. Obviously the better accuracy of the translation and intelligibility of c are the better quality of the translation should be expected and, on the contrary, the worse intelligibility of s is the worse quality of translation should be expected.

6 Future work

We plan to implement all of mentioned features into a new method of MT evaluation with working name LAMENT (LAnguage model and Meaning based Evaluation of machiNe Translation). The method should prove or falsify a hidden hypothesis: if it is possible to divide MT evaluation into two parts – to separate checking of intelligibility of a candidate translation (with help of language models) from matching words between a source sentence and a candidate translation. And, at the same time, provide a sufficient correlation with human evaluation not requiring any referential translations.

7 Conclusion

We have concisely described several commonly used methods of MT evaluation and commented their problems and shortcomings. Since these methods use manually prepared referential translations they could be regarded as semi-automatic methods. They simplified a process of developing new MT systems remarkably: if we include a new rule into our MT system we can instantly check out its impact on performance of the system. Despite these benefits they aren't suitable for an evaluation of arbitrary translations since they aren't versatile.

It may seem, after having mentioned the paradox of an evaluation, that developing of an universal MT evaluation methods is waste of time. But MT systems and MT evaluation methods are strongly interconnected therefore thinking of these methods helps us also with understanding of machine translation and with understanding of natural language in general. That is why it is worth dealing with them.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009.

References

1. Papineni, K., Roukos, S., Ward, T., and Zhu, W.: *A method for automatic evaluation of machine translation.* In: Proceedings of the 40th Annual Meeting on Association For Computational Linguistics. 2002.
2. Forbsom, E.: *Training a Super Model Look-Alike: Featuring Edit Distance, N-Gram Occurrence, and One Reference Translation.* In Proceedings of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation. 2003.
3. Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J.: *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. Cambridge. 2006.
4. Satanjeev B.: *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005.
5. Culy C., Riehemann S. Z.: *The limits of N-gram translation evaluation metrics*. Machine Translation Summit IX. 2003.