Manual and Automatic Paraphrases for MT Evaluation

Aleš Tamchyna, Petra Barančíková
Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
{tamchyna,barancikova}@ufal.mff.cuni.cz

Abstract

Paraphrasing of reference translations has been shown to improve the correlation with human judgements in automatic evaluation of machine translation (MT) outputs. In this work, we present a new dataset for evaluating English-Czech translation based on automatic paraphrases. We compare this dataset with an existing set of manually created paraphrases and find that even automatic paraphrases can improve MT evaluation. We have also propose and evaluate several criteria for selecting suitable reference translations from a larger set.

Keywords: machine translation, automatic evaluation, paraphrasing

1. Introduction

In a language, an idea might be expressed in an immense number of ways. This poses a serious issue for many linguistic tasks and particularly for machine translation (MT) evaluation.

In automatic machine translation evaluation, outputs of an MT system are compared to a reference translation, i.e. translation provided by a human translator. Due to the excessive amount of possible translations, metrics for this automatic comparison tend not to reflect human judgement very well.

These metrics incline to perform better if there are more reference sentences available. The pioneer metric BLEU (Papineni et al., 2002) was originally designed to work with four reference sentences.

However, obtaining reference sentences is labour intensive[1] and expensive. MT competitions, such as the Workshop on Statistical Machine Translation (WMT), use only one reference sentence for the automatic evaluation. This is still a considerable financial amount[2] given the number of languages and sentences.

Our motivation in this paper is to build on the work of human translators and to generate more reference translations automatically by paraphrasing the original manual translations. Our overall goal is to provide more robust evaluation of MT systems in this way.

There is evidence that paraphrasing of reference translations improves MT evaluation: in (Bojar et al., 2013b), human annotators were asked to create "all possible" paraphrases of 50 sentences from the WMT11 evaluation dataset (Callison-Burch et al., 2011). The authors show that correlation of MT evaluation metrics with human judgement increases with adding reference paraphrases.

In this work, we propose a simple approach to creating such a dataset automatically and compare the usefulness of automatic and manual paraphrases for MT evaluation.

Once such a large set of reference translations is created, it can simply be utilized within a standard metric such as BLEU. However, the number of possible paraphrased translations for a sentence can often be very high, which may result in unnecessarily expensive computation of system scores. In our work, we therefore also propose and evaluate several criteria for selecting a smaller sample of the reference sentences from the full set of paraphrased translations.

2. Related Work

Our work is closely related to (Bojar et al., 2013b). In this paper, annotators created up to several million paraphrases per reference sentence – Deprefset. While their work was relatively very efficient thanks to the possibility to formulate terse and expressive grammatical constraints in the annotation environment, the experiment was still costly in terms of human labour – annotators spent two hours per a sentence. Nonetheless, they managed to successfully increase the correlation with human judgements by adding these manual paraphrases as additional references.

Bloodgood and Callison-Burch (2010) shows that the cost of creating new references might be significantly lowered by using on-line crowd-sourcing services such as Amazon Mechanical Turk.

Positive effects of paraphrases on automatic metrics for MT evaluation are shown e.g. in Kauchak and Barzilay (2006), Owczarzak et al. (2006) or Barančíková et al. (2014).

Similarly, large paraphrase collections are employed to increase the quality of evaluation in several metrics, e.g. Meteor (Denkowski and Lavie, 2014), TERp (Snover et al., 2009) or ParaEval (Zhou et al., 2006). However, out of these metrics, only Meteor is available for MT evaluation of Czech; we use it for comparison with our results.

In Barančíková (2014), we show that evaluation using Meteor with exact match only on previously paraphrased reference sentences might lead to much higher correlation than using Meteor with paraphrase support. Therefore, we evaluate also using Meteor with no paraphrase support (MeteorNP).

[1] E.g. at the Linguistic Data Consortium, reference translation production is a complicated process which involves professional translation agencies, elaborate guidelines and thorough quality control (Strassel et al., 2006).

[2] For example, in the year 2009, the total cost of creating the test set was approximately $39,800 USD. It consists of 3027 sentences in seven European languages, which corresponds to slightly more than $0.08 USD/word (Bloodgood and Callison-Burch, 2010).

	systems	sentences	official score
WMT11	14/10	3003	">= others"
WMT13	14/12	3000	*Expected Wins*
WMT14	10	3003	*TrueSkill*

Table 1: Overview of WMT datasets. Number of systems translating from English to Czech (all MT systems / systems that were manually evaluated), number of source sentences and the official method for computing the absolute human judgement score.

3. Data

We use data sets from the English-to-Czech Translation Task of WMT from the years 2011, 2013 and 2014 (Callison-Burch et al., 2011; Bojar et al., 2013a; Bojar et al., 2014). Our main experiments are carried out with the WMT11 dataset, which we choose for comparability as it contains the 50 sentences that Deprefset was created from. Data from WMT13 and WMT14 are only used for validating the most promising approach.

All of these datasets consist of files with (Czech) outputs of MT systems, one file with corresponding reference sentences and one file with the original English source sentences. They differ in the number of MT systems and the length of the source files (see Table 1). We perform morphological analysis and tagging of the MT outputs and the reference sentences using Morphodita (Straková et al., 2014).

During the manual evaluation of WMT competitions, human judges fluent in both the source and the target language scores five MT outputs from the best to the worst translation. Thus, the human evaluation of MT system outputs is available as a relative ranking of performance of five systems for a sentence.

There are many ways to compute the absolute system score from this relative ranking. The official methods for each year are presented in Table 1 and we refer to these as the *gold standard*. The official method is different for every year. Therefore, to make our evaluation internally consistent, we also compute another absolute score for every year using the "> others" method (Bojar et al., 2011). This score is computed simply as $\frac{wins}{wins+loses}$ (it disregards ties). We refer to this interpretation of the human judgments as *silver standard* to distinguish it from the official system scores.

3.1. Sources of Paraphrases

We use all the paraphrase tables available for Czech: Czech WordNet (Pala and Smrž, 2004), Meteor Paraphrase Tables (Denkowski and Lavie, 2010) and The Paraphrase Database (PPDB) (Ganitkevitch and Callison-Burch, 2014).

We use the first two tables for lexical (one-word) paraphrasing. We attempt to reduce the noise in the Meteor paraphrase tables by additional filtering based on POS. We do not use multi-word paraphrases from the Czech Meteor paraphrase tables as Barančíková et al. (2014) shows that they are so noisy that they actually harm the results of metrics applied on paraphrased sentences.

For multi-word paraphrases, we thus apply PPDB only. We combine the following tables: phrasal paraphrases, many-to-one and one-to-to many. PPDB offers several sizes of paraphrase tables, ranging from S to XXL. We choose the size L for all of them as an optimal trade-off between size and noise.

3.2. Deprefset

Deprefset (Bojar et al., 2013b) was created by human annotators using two interfaces (Prolog command-line interface and a web interface). The authors dub their approach "unification-based annotation". Annotators propose alternative phrasing of various parts of the sentences and create ad-hoc unification constraints to ensure grammatical consistency of the generated paraphrases. All combinations of the proposed variants which conform to the constraints are then automatically generated, constituting the final Deprefset.

4. Paraphrase Generation

We use a simple brute-force algorithm to create the paraphrases. For every reference translation $T = t_1, ..., t_n$, there are available several MT outputs $H_i = h_{i,1}, ..., h_{i,m_i}$.

For every word t_x, we create a set of paraphrases P_{t_x} which consists of words from the translations $h_{i,y}$ such that t_x and $h_{i,y}$ have the same POS and their lemmas are paraphrases according to WordNet or filtered Meteor tables.

Once we have collected the possible paraphrases for words from the reference translation, we proceed to create the paraphrased reference sentences: we step word by word from the beginning of the reference sentence, creating partial sentences from all original words and paraphrastic matches: $Ref_0 = \emptyset$ to $Ref_n = Ref_{n-1} \times \{t_n \cup P_{t_x}\}$.

For multi-word paraphrases, we first extract all pairs of phrases up to 4 words from T and all H_i. We then filter them and keep only phrase pairs which are paraphrases according to PPDB. Next, we apply these multi-word paraphrases on the full sentences Ref_n (which have so far been paraphrased only word by word). We go through Ref_n and for every sentence with a multi-word paraphrase in, we add another one where the paraphrase match is substituted.

Due to memory limitations, we put a restriction on the amount of generated paraphrases. When the number of possible paraphrases for a sentence grows too large, we shuffle all generated paraphrases randomly and pick only the first 10 000.

We provide the automatic paraphrases of WMT11, WMT13 and WMT14 test sets as a freely available dataset at the following URL:

`http://hdl.handle.net/11234/1-1665`

5. Selection of Reference Sentences

For both the manual and automatic paraphrases, we have a large set of possible reference translations which could be used. We use R to denote the full set of references (the maximum size of R is 10000, see previous section). Our goal is to use as few reference translations as possible, so that the effort in creating such a dataset is minimized and that the computation of system scores does not become unnecessarily expensive.

	Deprefset			Paraphrases – 50			Paraphrases – Full		
	1	10	1000	1	10	1000	1	10	1000
bestLM	0.53	0.62	0.70	0.57	0.54	0.60	0.61	0.63	0.67
worstLM	0.86	0.86	0.73	0.70	0.70	0.63	0.83	0.82	0.72
dissimilar	0.69	0.70	–	0.75	0.60	–	0.73	0.66	–
random	0.71±0.06	0.70±0.02	0.73±0.00	0.71±0.06	0.61±0.03	0.60±0.00	0.69±0.01	0.67±0.00	0.66±0.00

Table 2: Correlations of BLEU with gold standard for the WMT11 evaluation set. Correlation of the official evaluation set is 0.65.

We evaluate several criteria for selecting references from R. For each criterion, we vary the number of selected sentences k. We use the subsets R_k for the evaluation of MT systems with BLEU and then measure the correlation with human judgements.

Our results can be used to optimize the effort when creating sets of reference translations for MT. They provide a way to estimate the benefit of creating more references (i.e., to judge how returns diminish as k grows) and they might serve as a guide when creating additional references. In the following paragraphs, we describe the evaluated criteria.

5.1. Random Selection

As a baseline, we simply shuffle all the generated reference translations R and use the top k sentences as R_k.

5.2. Language Model Perplexity

We use a language model (LM) trained on 1 million Czech sentences from the CzEng 1.0 corpus (Bojar et al., 2012). We calculate the LM perplexity for each reference translation in R. We then select either top k (lowest perplexity – most "fluent" sentences) or bottom k (highest perplexity) as the R_k.

5.3. Diversity

It is intuitive to aim for a set of references which is as diverse as possible. We could expect that such an approach to reference selection can provide a representative sample of the full reference set R even when k is small.

Our measure of similarity is the Levenshtein distance. We construct the set of dissimilar references as follows: we begin with the official reference translation[3] and search R for the most dissimilar sentence. In the subsequent steps, we go over all sentences which remain in R and look for the one which is on average the least similar to the sentences already selected. We continue adding the least similar sentence from R to our solution until we select k references.

While this approach does not construct the optimal most diverse R_k, we believe that it is a sufficient approximation for this evaluation.

6. Experiments

We evaluate our automatically generated paraphrases in two settings – either only using the same 50 sentences as Deprefset or using the full test set (3003 sentences). In the first setting, our results are directly comparable with Deprefset.

[3]The official reference translation is later removed from the set.

For each setting, we compute BLEU for all systems using a single reference or large numbers references, depending on the setting. We then measure the correlation of these BLEU scores with manual ranking.

Table 2 shows the results of Deprefset and our paraphrased references in the two settings. For selection criteria, note that most of them are deterministic, except for random selection where we therefore also report standard deviation.

6.1. Automatic and Manual Paraphrases

Our aim in this section is to compare the automatic and manual paraphrases (in terms of correlation with human judgement). We therefor look at the random selection: if we use 1, 10, or 1000 random sentences from the given set, what correlation can we expect?

When selecting just a single sentence, all approaches look similar with the correlation around 0.70. Considering that the correlation of the official references is only 0.65, random selection (both its mean and standard deviation) provides an interesting perspective: we could view the official dataset as a slightly unfortunate sample from the possible paraphrases.

When we use 10 or 1000 sentences, the difference becomes clear – manual paraphrases provide a better reference set than the automatic ones. When we compare Deprefset with our paraphrases on the same data, i.e. only 50 sentences, the difference in correlation is over 0.1 absolute (0.73 vs. 0.60). Randomly selecting from the automatic paraphrases is no better than using the official reference translations.

It is apparent that in order for the automatic paraphrases to be useful, we need a better way of selecting a good subset from them. Indeed, when we use a suitable criterion for reference selection, we can obtain a correlation of 0.83 with the automatic paraphrases (as opposed to the 0.65 achieved with the official reference translations).

6.2. Selection Criteria

There does not seem to be much difference between selecting the sentences randomly and using the diversity criterion. Both methods apparently quickly create a representative sample of the full set.

Criteria based on LM perplexity show interesting results: using the most fluent sentences from R leads to the worst correlation while using the worst sentences (highest perplexity according to the LM) is superior to all the evaluated methods. Interestingly, the smaller set the better with the WorstLM criterion and adding more fluent sentence harms the performance of BLEU. In fact, by using this criterion, we are able to obtain a far better correlation with human judgements than by using full set of paraphrases.

original reference	*cigarety*	*stojí*	*za*	*85*	*%*	*případů*	*rakoviny*	*plic*	.
	cigarettes	cause	for	85	%	cases	cancer	lungs	.
	cigarettes cause 85% of lung cancer cases.								
worstLM	*cigarety*	*spojeny*	*za*	*85*	*%*	*případů*	*karcinomu*	*plic*	.
	cigarettes	connected	for	85	%	cases	carcinoma	lungs	.
	*cigarettes are connected with 85% of lung carcinoma cases.								

MT system	output										human	reference	worstLM
commercial-1	cigarety	jsou	spojené	s	85	%	rakoviny	plic	případů	.	0.11	0.37	0.29
	cigarettes	are	connected$_{adj}$	with	85	%	cancer	lungs	cases	.			
	*cigarettes are connected with 85% of lung cancer cases.												
cu-zeman	cigarety	jsou	napojeny	na	85	%	případů	rakoviny	plic	.	0.17	0.44	0.33
	cigarettes	are	attached	to	85	%	cases	cancer	lungs	.			
	*cigarettes are attached to 85% of lung cancer cases.												
cu-bojar	cigarety	jsou	spojeny	s	85	%	případů	rakoviny	plic	.	1.00	0.44	0.39
	cigarettes	are	connected$_{verb}$	with	85	%	cases	lung	cancer	.			
	cigarettes are connected with 85% of lung cancer cases.												
											correlation	0.64	0.95

Table 3: An example sentence from the WMT13 test set. *Human* stands for human judgement computed using the silver standard method on this sentence only, *reference* and *worstLM* shows MeteorNP scores computed on the original reference and the worstLM paraphrase, respectively. Other systems are skipped – they produced one of the already presented sentences.

WMT11

metric	dataset	gold stand.	silver stand.
Meteor	official	0.63	0.53
BLEU		0.65	0.53
	worstLM	0.83	0.74
MeteorNP		**0.84**	**0.78**

WMT13

metric	dataset	gold stand.	silver stand.
Meteor	official	0.83	0.82
BLEU		0.84	0.84
	worstLM	0.92	0.92
MeteorNP		**0.95**	**0.95**

WMT14

metric	dataset	gold stand.	silver stand.
Meteor	official	0.96	0.97
BLEU		0.97	0.97
	worstLM	0.96	0.97
MeteorNP		0.96	0.97

Table 4: Correlations with human rankings of the official reference translations and of paraphrased translations selected based on the LM perplexity.

We can observe a similar effect happens also on the other side of LM scale – increasing size of set of references with best LM score (i.e adding less fluent sentences) helps to increase correlation (moving it closer to the "average").

To validate further validate our results, we create paraphrases also for WMT13 and WMT14 and compare the official reference set to a set of paraphrased references; for each sentence, we only use a single paraphrase and we select the one with highest LM perplexity.[4] Table 4 shows that our results are consistent.

For both the silver and the gold standard, the baseline correlation is improved significantly for WMT11 and WMT13. It is only negligibly worsened for WMT14, measured by the gold standard (where the baseline already has close-to-perfect correlation with human ranking).

Metrics such as BLEU tend to reward fluency more than adequacy (Wu and Fung, 2009). Most of the evaluated systems are statistical and they are optimized towards one of the automatic metrics. This can lead them to rely on LMs heavily and prefer fluency to adequacy.

By selecting the least "fluent" sentence from the set of possible translations, we reduce the benefit for BLEU that the systems get from using a strong LM. When we use these disfluent reference translations in the automatic metric, the differences in translation adequacy become prominent and possibly bring the automatic scores closer to the human judgements.

An example of our method is presented in Table 3. The paraphrase selected by the WorstLM criterion contains the word "spojeny" ("connected", a paraphrase of "caused") which allows to correctly distinguish between the systems `cu-zeman` and `cu-bojar`.

6.3. Automatic Metrics

Both BLEU and Meteor behave consistently. We obtained the worst results by using Meteor with paraphrase support on official reference sentences. The main reason is the noise in its paraphrase tables. The metric may award even parts of the hypothesis which are left untranslated, as the Czech Meteor paraphrase tables contain even English words and their Czech translations as paraphrases, for example: *pšenice - wheat*,[5] *vůdce - leader*, *vařit - cook*.

On a WorstLM reference, results of BLEU and MeteorNP are similar with MeteorNP having slight upper hand as BLEU with its strong emphasis on longer n-gram is not best suited for evaluating languages with free word order such as Czech.

[4] Deprefset is not available for these datasets.

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

7. Conclusion

We have presented a new dataset for evaluating English-Czech translation based on automatic paraphrases. We have also proposed several criteria for selecting suitable reference translations from a larger set.

The best way of selecting reference translations in our setting is to choose the references with the highest perplexity according to a language model trained on monolingual data in the target language. By choosing references which are "disfluent", differences in translation adequacy become more prominent and this improves the automatic ranking.

Acknowledgement

This research was supported by the grants H2020-ICT-2014-1-645452 (QT21), GAUK 1356213 and by SVV project number 260224. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

8. Bibliographical References

Barančíková, P., Rosa, R., and Tamchyna, A. (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.

Barančíková, P. (2014). Parmesan: Meteor without Paraphrases with Paraphrased References. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 355–361, Baltimore, MD, USA. Association for Computational Linguistics.

Bloodgood, M. and Callison-Burch, C. (2010). Using mechanical turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 208–211, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. F. (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.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Kauchak, D. and Barzilay, R. (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.

Owczarzak, K., Groves, D., Van Genabith, J., and Way, A. (2006). Contextual bitext-derived paraphrases in automatic mt evaluation. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 86–93, Stroudsburg, PA, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Snover, M. G., Madnani, N., Dorr, B., and Schwartz, R. (2009). TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3):117–127, September.

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics.

Strassel, S., Cieri, C., Cole, A., Dipersio, D., Liberman, M., Maamouri, M., and Maeda, K. (2006). Integrated linguistic resources for language exploitation technologies. In *In Proceedings of LREC*.

Wu, D. and Fung, P. (2009). Semantic roles for smt: A hybrid two-pass model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 13–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhou, L., yew Lin, C., and Hovy, E. (2006). Re-evaluating Machine Translation Results with Paraphrase Support. In *In Proceedings of EMNLP*.

9. Language Resource References

Ondřej Bojar and Zdeněk Žabokrtský and Ondřej Dušek and Petra Galuščáková and Martin Majliš and David Mareček and Jiří Maršík and Michal Novák and Martin Popel and Aleš Tamchyna. (2012). *The Joy of Parallelism with CzEng 1.0*. European Language Resources Association.

Ondřej Bojar and Christian Buck and Chris Callison-Burch and Christian Federmann and Barry Haddow and Philipp Koehn and Christof Monz and Matt Post and Radu Soricut and Lucia Specia. (2013a). *Findings of the 2013 Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Ondřej Bojar and Matouš Macháček and Aleš Tamchyna and Daniel Zeman. (2013b). *Scratching the Surface of Possible Translations*. Springer Verlag.

Ondřej Bojar and Christian Buck and Christian Federmann and Barry Haddow and Philipp Koehn and Leveling, Johannes and Christof Monz and Pavel Pecina and Matt Post and Herve Saint-Amand and Radu Soricut and Lucia Specia and Aleš Tamchyna. (2014). *Findings of the 2014 Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Chris Callison-Burch and Philipp Koehn and Christof Monz and Omar Zaidan. (2011). *Findings of the 2011*

Workshop on Statistical Machine Translation. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. (2010). *METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages*.

Juri Ganitkevitch and Chris Callison-Burch. (2014). *The Multilingual Paraphrase Database*. European Language Resources Association.

Karel Pala and Pavel Smrž. (2004). *Building Czech WordNet*.