Targeted Paraphrasing on Deep Syntactic Layer for MT Evaluation

Petra Barančíková and **Rudolf Rosa**
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
{barancikova,rosa}@ufal.mff.cuni.cz

Abstract

In this paper, we present a method of improving quality of machine translation (MT) evaluation of Czech sentences via targeted paraphrasing of reference sentences on a deep syntactic layer. For this purpose, we employ NLP framework Treex and extend it with modules for targeted paraphrasing and word order changes. Automatic scores computed using these paraphrased reference sentences show higher correlation with human judgment than scores computed on the original reference sentences.

1 Introduction

Since the very first appearance of machine translation (MT) systems, a necessity for their objective evaluation and comparison has emerged. The traditional human evaluation is slow and unreproducible; thus, it cannot be used for tasks like tuning and development of MT systems. Well-performing automatic MT evaluation metrics are essential precisely for these tasks.

The pioneer metrics correlating well with human judgment were BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). They are computed from an n-gram overlap between the translated sentence (hypothesis) and one or more corresponding reference sentences, i.e., translations made by a human translator.

Due to its simplicity and language independence, BLEU still remains the de facto standard metric for MT evaluation and tuning, even though other, better-performing metrics exist (Macháček and Bojar (2013), Bojar et al. (2014)).

Furthermore, the standard practice is using only one reference sentence and BLEU then tends to perform badly. There are many translations of a single sentence and even a perfectly correct translation might get a low score as BLEU disregards synonymous expressions and word order variants (see Figure 1). This is especially valid for morphologically rich languages with free word order like the Czech language (Bojar et al., 2010).

In this paper, we use deep syntactic layer for targeted paraphrasing of reference sentences. For every hypothesis, we create its own reference sentence that is more similar in wording but keeps the meaning and grammatical correctness of the original reference sentence. Using these new paraphrased references makes the MT evaluation metrics more reliable. In addition, correct paraphrases have additional application in many other NLP tasks.

As far as we know, this is the first rule-based model specifically designed for targeted paraphrased reference sentence generation to improve MT evaluation quality.

2 Related Work

Second generation metrics Meteor (Denkowski and Lavie, 2014), TERp (Snover et al., 2009) and ParaEval (Zhou et al., 2006) still largely focus on an n-gram overlap while including other linguistically motivated resources. They utilize paraphrase support in form of their own paraphrase tables (i.e. collection of synonymous expressions) and show higher correlation with human judgment than BLEU.

Meteor supports several languages including Czech. However, its Czech paraphrase tables are so noisy (i.e. they contain pairs of non-paraphrastic expressions) that they actually harm the performance of the metric, as it can reward mistranslated and even untranslated words (Barančíková, 2014).

String matching is hardly discriminative enough to reflect the human perception and there is growing number of metrics that compute their score based on rich linguistic features and matching based on parse trees, POS tagging or textual entail-

Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 20–27,
Uppsala, Sweden, August 24–26 2015.

Original sentence	*Banks are testing payment by mobile telephone*					
Hypothesis	*Banky*	*zkoušejí*	*platbu*	*pomocí*	*mobilního*	*telefonu*
	Banks	are testing	payment	with help	mobile	phone
	Banks are testing payment by mobile phone					
Reference sentence	*Banky*	*testují*	*placení*	*mobilem*		
	Banks	are testing	paying	by mobile phone		
	Banks are testing paying by mobile phone					

Figure 1: Example from WMT12 - Even though the hypothesis is grammatically correct and the meaning of both sentences is the same, it doesn't contribute to the BLEU score. There is only one unigram overlapping.

ment (e.g. Liu and Gildea (2005), Owczarzak et al. (2007), Amigó et al. (2009), Padó et al. (2009), Macháček and Bojar (2011)).

These metrics shows better correlation with human judgment, but their wide usage is limited by being complex and language-dependent. As a result, there is a trade-off between linguistic-rich strategy for better performance and applicability of simple string level matching.

Our approach makes use of linguistic tools for creating new reference sentences. The advantage of this method is that we can choose among many traditional metrics for evaluation on our new references while eliminating some shortcomings of these metrics.

Targeted paraphrasing for MT evaluation was introduced by Kauchak and Barzilay (2006). Their algorithm creates new reference sentences by one-word substitution based on WordNet (Miller, 1995) synonymy and contextual evaluation. This solution is not readily applicable to the Czech language – a Czech word has typically many forms and the correct form depends heavily on its context, e.g., morphological cases of nouns depend on verb valency frames. Changing a single word may result in an ungrammatical sentence. Therefore, we do not attempt to change a single word in a reference sentence but we focus on creating one single correct reference sentence.

In Barančíková and Tamchyna (2014), we experimented with targeted paraphrasing using the freely available SMT system Moses (Koehn et al., 2007). We adapted Moses for targeted monolingual phrase-based translation. However, results of this method was inconclusive. It was mainly due to a high amount of noise in the translation tables and unbalanced targeting feature.

As a result, we rather chose to employ rule-based translation system. This approach has many advantages, e.g. there is no need for creating a targeting feature and we can change only parts of a sentence and thus create more conservative paraphrases. We utilize Treex (Popel and Žabokrtský, 2010), highly modular NLP software system developed for machine translation system TectoMT (Žabokrtský et al., 2008) that translates on a deep syntactic layer. We performed our experiment on the Czech language, however, we plan to extend it to more languages, including English and Spanish.

Treex is open-source and is available on GitHub,[1] including the two blocks that we contributed. In the rest of the paper, we describe the implementation of our approach.

3 Treex

Treex implements a stratificational approach to language, adopted from the Functional Generative Description theory (Sgall, 1967) and its later extension by the Prague Dependency Treebank (Bejček et al., 2013). It represents sentences at four layers:

- **w-layer:** word layer; no linguistic annotation
- **m-layer:** morphological layer; sequence of tagged and lemmatized tokens
- **a-layer:** shallow-syntax/analytical layer; sentence is represented as a surface syntactic dependency tree
- **t-layer:** deep-syntax/tectogrammatical layer; sentence is represented as a deep-syntactic dependency tree, where autosemantic words (i.e. semantically full lexical units) only have their own nodes; t-nodes consist of a t-lemma and a set of attributes – a *formeme* (information about the original syntactic form) and a

[1] `https://github.com/ufal/treex`

Source	*The Internet has caused a boom in these speculations.*						
Hypothesis	Internet	vyvolal	boom	v	těchto	spekulacích	.
	Internet	*caused*	*boom*	*in*	*these*	*speculations*	.
	The Internet has caused a boom in these speculations.						
Reference	Rozkvět	těchto	spekulací	způsobil	internet	.	
	Boom	*these*	*speculations*	*caused*	*internet*	.	
	A boom of these speculation was caused by the Internet.						

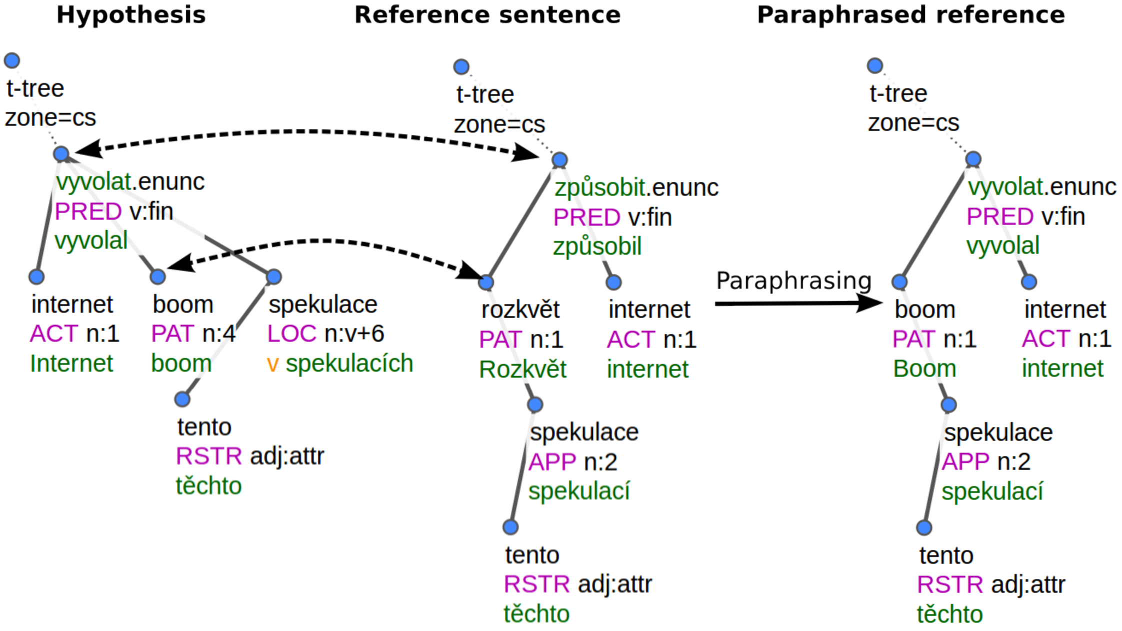

Figure 2: Example of the paraphrasing. The hypothesis is grammatically correct and has the same meaning as the reference sentence. We analyse both sentences to t-layer, where we create a new reference sentence by substituting synonyms from hypothesis to the reference. In the next step, we will change also the word order to better reflect the hypothesis.

set of *grammatemes* (essential morphological features).

We take the analysis and generation pipeline from the TectoTM system. We transfer both a hypothesis and its corresponding reference sentence to the t-layer, where we integrate a module for t-lemma paraphrasing. After paraphrasing, we perform synthesis to a-layer, where we plug in a reordering module and continue with synthesis to the w-layer.

3.1 Analysis from w-layer to t-layer

The analysis from the w-layer the to a-layer includes tokenization, POS-tagging and lemmatization using MorphoDiTa (Straková et al., 2014), dependency parsing using the MSTParser (McDonald et al., 2005) adapted by Novák and Žabokrtský (2007), trained on PDT.

In the next step, a surface-syntax a-tree is converted into a deep-syntax t-tree. Auxiliary words are removed, with their function now represented using t-node attributes (grammatemes and formemes) of autosemantic words that they belong to (e.g. two a-nodes of the verb form *spal jsem* (“I slept”) would be collapsed into one t-node *spát* (“sleep”) with the tense grammateme set to past; *v květnu* (“in May”) would be collapsed into *květen* (“May”) with the formeme *v+X* (“in+X”).

We choose the t-layer for paraphrasing, because the words from the sentence are lemmatized and free of syntactical information. Furthermore, functional words, which we do not want to paraphrase and that cause a lot of noise in our paraphrase tables, do not appear here.

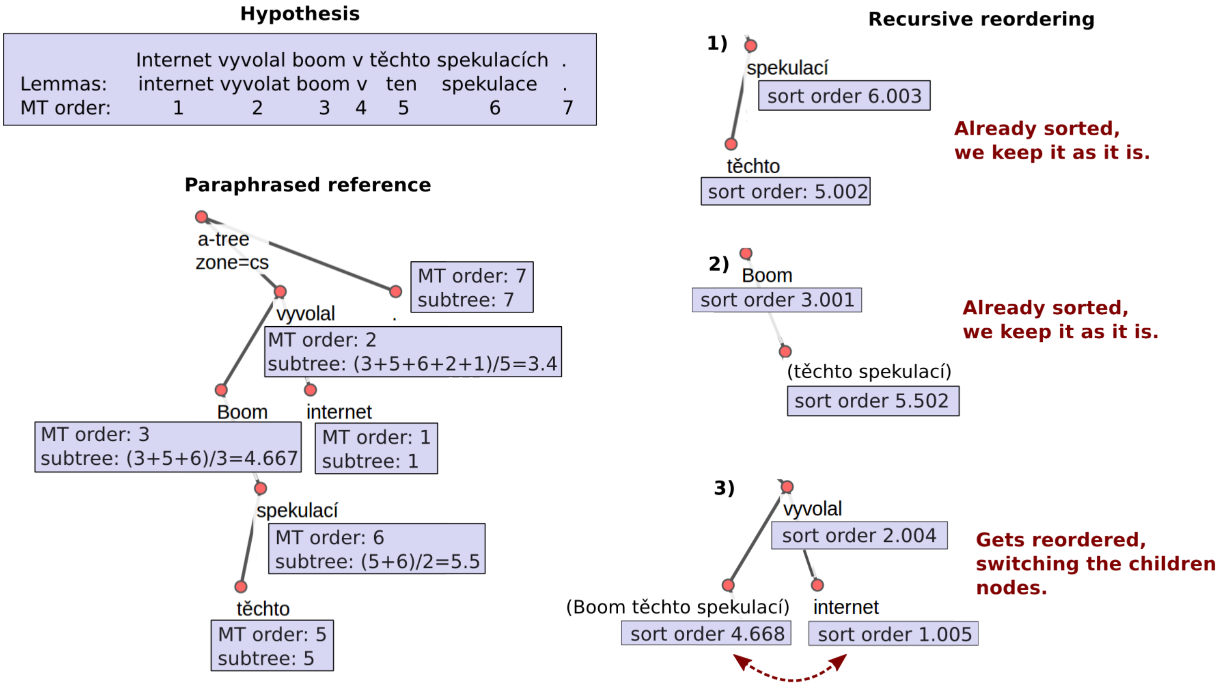

Figure 3: Continuation of Figure 2, reordering of the paraphrased reference sentence.

3.2 Paraphrasing

The paraphrasing module T2T::ParaphraseSimple is freely available at GitHub.[2]

T-lemma of a reference t-node R is changed from A to B if and only if:

1. there is a hypothesis t-node with lemma B
2. there is no hypothesis t-node with lemma A
3. there is no reference t-node with lemma B
4. A and B are paraphrases according to our paraphrase tables

The other attributes of the t-node are kept unchanged based on the assumption that semantic properties are independent of the t-lemma. However, in practice, there is at least one case where this is not true: t-nodes corresponding to nouns are marked for grammatical gender, which is very often a grammatical property of the given lemma with no effect on the meaning (for example, “a house” can be translated either as a masculine noun *dům* or as feminine noun *budova*),

Therefore, when paraphrasing a t-node that corresponds to a noun, we delete the value of the gender grammateme, and let the subsequent synthesis pipeline generate the correct value of the morphological gender feature value (which is necessary to ensure correct morphological agreement of the noun’s dependents, such as adjectives and verbs).

3.3 Synthesis from t-layer to a-layer

In this phase, a-nodes corresponding to auxiliary words and punctuation are generated, morphological feature values on a-nodes are initialized and set to enforce morphological agreement among the nodes. Correct inflectional forms based on lemma and POS, and morphological features are generated using MorphoDiTa.

3.4 Tree-based reordering

The reordering block A2A::ReorderByLemmas is freely available at GitHub.[3]

The idea behind the block is to make the word order of the new reference as similar to the word order of the translation, but with some tree-based constraints to avoid ungrammatical sentences.

The general approach is to reorder the subtrees rooted at modifier nodes of a given head node so that they appear in an order that is on average similar to their order in the translation. Figure 3 shows the reordering process of the a-tree from Figure 2.

[2] https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/ParaphraseSimple.pm

[3] https://github.com/ufal/treex/blob/master/lib/Treex/Block/A2A/ReorderByLemmas.pm

Our reordering proceeds in several steps. Each a-node has an order, i.e. a position in the sentence. We define the *MT order* of a reference a-node as the order of its corresponding hypothesis a-node, i.e. a node with the same lemma.

We set the MT order only if there is exactly one a-node with the given lemma in both the hypothesis and the reference. Therefore, the MT order might be undefined for some nodes.

In the next step, we compute the *subtree MT order* of each reference a-node R as the average MT order of all a-nodes in the subtree rooted at the a-node R (including the MT order of R itself). Only nodes with a defined MT order are taken into account, so the subtree MT order can be undefined for some nodes.

Finally, we iterate over all a-nodes recursively starting from the bottom. Head a-node H and its dependent a-nodes D_i are reordered if they violate the *sorting order*. If D_i is a root of a subtree, the whole subtree is moved and its internal ordering is kept.

The sorting order of H is defined as its MT order; the sorting order of each dependent node D_i is defined as its subtree MT order. If a sorting order of a node is undefined, it is set to the sorting order of the node that precedes it, thus favouring neighbouring nodes (or subtrees) to be reordered together in case there is no evidence that they should be brought apart from each other. Additionally, each sorting order is added 1/1000th of the original order of the node – in case of a tie, the original ordering of the nodes is preferred to reordering.

We do not handle non-projective edges in any special way, so they always get projectivized if they take part in a reordering process, or kept in their original order otherwise. However, no new non-projective edges are created in the process – this is ensured by always moving the subtrees at once.

Please note that each node can take part in at most two reorderings – once as the H node and once as a D_i node. Moreover, the nodes can be processed in any order, as a reordering does not influence any other reordering.

3.5 Synthesis from a-layer to w-layer

The word forms are already generated on the a-layer, so there is little to be done. Superfluous tokens are deleted (e.g. duplicated commas)the first letter in a sentence is capitalized, and the tokens are concatenated (a set of rules is used to decide which tokens should be space-delimited and which should not). The example in Figure 3) results in the following sentence: *Internet vyvolal boom těchto spekulací* ("The Internet has caused a boom of these speculations."), which has the same meaning as the original reference sentence, is grammatically correst and, most importantly, is much more similar in wording to the hypothesis.

4 Data

We perform our experiments on data sets from the English-to-Czech translation task of WMT12 (Callison-Burch et al., 2012), WMT13 (Bojar et al., 2013a). The data sets contain 13/14[4] files with Czech outputs of MT systems. Each data set also contains one file with corresponding reference sentences.

Our database of t-lemma paraphrases was created from two existing sources of Czech paraphrases – the Czech WordNet 1.9 PDT (Pala and Smrž, 2004) and the Meteor Paraphrase Tables (Denkowski and Lavie, 2010). Czech WordNet 1.9 PDT is already lemmatized, lemmatization of the Meteor Paraphrase tables was performed using MorphoDiTa (Straková et al., 2014).

We also performed fitering of the lemmatized Meteor Paraphrase tables based on coarse POS, as they contained a lot of noise due to being constructed automatically.

5 Results

The performance of an evaluation metric in MT is usually computed as the Pearson correlation between the automatic metric and human judgment (Papineni et al., 2002). The correlation estimates the linear dependency between two sets of values. It ranges from -1 (perfect negative linear relationship) to 1 (perfect linear correlation).

The official manual evaluation metric of WMT12 and WMT13 provides just a relative ranking: a human judge always compares the performance of five systems on a particular sentence. From these relative rankings, we compute the absolute performance of every system using the "> others" method (Bojar et al., 2011). It is computed as $\frac{wins}{wins+loses}$.

Our method of paraphrasing is independent of an evaluation metric used. We employ three dif-

[4] We use only 12 of them because two of them (FDA.2878 and online-G) have no human judgments.

	WMT12			**WMT13**		
references	original	paraphrased	reordered	original	paraphrased	paraphrased
BLEU	0.751	0.783	0.804	0.834	0.850	0.878
Meteor	0.833	0.864	0.868	0.817	0.871	0.870
Ex.Meteor	0.861	0.900	**0.903**	0.848	**0.893**	**0.893**

Table 1: Pearson correlation of a metric and human judgment on original references, paraphrased references and paraphrased reordered references. Ex.Meteor represents Meteor metric with exact match only (i.e. no paraphrase support).

ferent metrics - BLEU score, Meteor metric and Meteor metric without the paraphrase support (as it seem redundant to use paraphrases on already paraphrased sentences).

The results are presented in Table 1 as a Pearson correlation of a metric with human judgment. Paraphrasing clearly helps to reflect the human perception better. Even the Meteor metric that already contains paraphrases is performing better using paraphrased references created from its own paraphrase table. This is again due to the noise in the paraphrase table, which blurs the difference between the hypotheses of different MT systems.

The reordering clearly helps when we evaluate via the BLEU metric, which punishes any word order changes to the reference sentence. Meteor is more tolerant to word order changes and the reordering has practically no effect on his scores.

However, manual examination showed that our constraints are not strong enough to prevent creating ungrammatical sentences. The algorithm tends to copy the word order of the hypothesis, even if it is not correct. Most errors were caused by changes of a word order of punctuation.

6 Future Work

In our future work, we plan to extend the paraphrasing module for more complex paraphrases including syntactical paraphrases, longer phrases, diatheses. We will also change only parts of sentences that are dependent on paraphrased words, thus keeping the rest of the sentence correct and creating more conservative reference sentences.

We also intend to adjust the reordering function by adding rule-based constrains. Furthermore, we'd like to learn automatically possible word order changes from Deprefset (Bojar et al., 2013b), which contains an excessive number of manually created reference translations for 50 Czech sentences.

We performed our experiment on Czech language, but the procedure is generally language independent, as long as there is analysis and synthesis support for particular language in Treex. Currently there is full support for Czech, English, Portuguese and Dutch, but there is ongoing work on many more languages within the QTLeap[5] project.

Acknowledgments

This research was supported by the following grants: SVV project number 260 224 and GAUK 1572314. This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

References

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Felisa Verdejo. 2009. The Contribution of Linguistic Features to Automatic Machine Translation Evaluation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 306–314.

Petra Barančíková. 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.

Petra Barančíková and Aleš Tamchyna. 2014. Machine Translation within One Language as a Paraphrasing Technique. In *Proceedings of the main track of the 14th Conference on Information Technologies - Applications and Theory (ITAT 2014)*, pages 1–6.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda

[5] `http://qtleap.eu/`

Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague Dependency Treebank 3.0.

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 86–91, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar F. Zaidan. 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.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013a. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. 2013b. Scratching the Surface of Possible Translations. In *Text, Speech and Dialogue: 16th International Conference, TSD 2013. Proceedings*, pages 465–474, Berlin / Heidelberg. Springer Verlag.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. In *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR*.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Cooccurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

David Kauchak and Regina Barzilay. 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.

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 *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ding Liu and Daniel Gildea. 2005. Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pages 25–32. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2011. Approximating a Deep-syntactic Metric for MT Evaluation and Tuning. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 92–98, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective Dependency Parsing Using Spanning Tree Algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530.

George A. Miller. 1995. WordNet: A Lexical Database for English. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Václav Novák and Zdeněk Žabokrtský. 2007. Feature Engineering in Maximum Spanning Tree Dependency Parser. In Václav Matousek and Pavel Mautner, editors, *TSD*, Lecture Notes in Computer Science, pages 92–98. Springer.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Labelled Dependencies in Machine Translation Evaluation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Padó, Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Measuring Machine Translation Quality as Semantic Equivalence: a Metric Based on Entailment Features. *Machine Translation*, 23(2-3):181–193, September.

Karel Pala and Pavel Smrž. 2004. Building Czech WordNet. *In Romanian Journal of Information Science and Technology*, 7:79–88.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, IceTAL'10, pages 293–304, Berlin, Heidelberg. Springer-Verlag.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Number v. 6 in Generativní popis jazyka a česká deklinace. Academia.

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

Jana Straková, Milan Straka, and Jan Hajič. 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.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. Tectomt: Highly modular mt system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 167–170.

Liang Zhou, Chin yew Lin, and Eduard Hovy. 2006. Reevaluating machine translation results with paraphrase support. In *In Proceedings of EMNLP*.