V. Kůrková et al. (Eds.): ITAT 2014 with selected papers from Znalosti 2014, CEUR Workshop Proceedings Vol. 1214, pp. 1–6
http://ceur-ws.org/Vol-1214, Series ISSN 1613-0073, © 2014 P. Barančíková, A. Tamchyna

Machine Translation within One Language as a Paraphrasing Technique

Petra Barančíková, Aleš Tamchyna

Institute of Formal and Applied Linguistics
Charles University in Prague, Faculty of Mathematics and Physics
Malostranské náměstí 25, Prague, Czech Republic
{barancikova,tamchyna}@ufal.mff.cuni.cz

Abstract: We present a method for improving machine translation (MT) evaluation by targeted paraphrasing of reference sentences. For this purpose, we employ MT systems themselves and adapt them for translating within a single language. We describe this attempt on two types of MT systems – phrase-based and rule-based. Initially, we experiment with the freely available SMT system Moses. We create translation models from two available sources of Czech paraphrases – Czech WordNet and the Meteor Paraphrase tables. We extended Moses by a new feature that makes the translation targeted. However, the results of this method are inconclusive. In the view of errors appearing in the new paraphrased sentences, we propose another solution – targeted paraphrasing using parts of a rule-based translation system included in the NLP framework Treex.

1 Introduction

In this paper, we examine the possibility of improving accuracy of metrics for automatic evaluation of MT systems by the machine translation itself.

The first metric correlating well with human judgment was BLEU [20] and it still remains the most common metric for MT evaluation, even though other, better-performing metrics exist. [15]

BLEU is computed from the number of phrase overlaps between the translated sentence and the corresponding reference sentences, i.e., translations made by a human translator. However, the standard practice is using only one reference sentence and BLEU then tends to perform badly.

As there are many translations of a single sentence, even a perfectly correct machine translation might get a low score due to disregarding synonyms and paraphrase expressions. This is especially valid for morphologically rich languages like the Czech language. [7]

We aim to achieve higher accuracy of MT evaluation by targeted paraphrasing of reference sentences, i.e. creating a new synthetic reference sentence that is still correct and keeps the meaning of the original sentence, but at the same time it is closer in wording to the MT output (hypothesis).

There is a close resemblance between translation and paraphrasing. They both attempt to preserve the meaning of a sentence, the first one between two languages and the second one within one language by different word choice. [16] However, there are many more tools for MT than for paraphrasing. Therefore, it seems only natural to attempt to adjust some MT tools to translate within a single language for targeted paraphrasing.

2 Related Work

In [3], a significant improvement in correlation of BLEU with human judgment was achieved by targeted paraphrasing of Czech reference sentences. However, the best results were acquired using a simple greedy algorithm for one-word paraphrase substitution, which does not allow word order changes and other alternation of reference sentence. The grammatical correctness was achieved by applying Depfix [22], an automatic post-editing system, originally designed for improving quality of phrase-based English-to-Czech machine translation outputs.

[13] used lexical substitution and contextual evaluation to improve the accuracy of Chinese-to-English MT evaluation. In [16], targeted paraphrasing via SMT is used to improve SMT itself during the parameter optimization phase of machine translation. Correct hypotheses are no longer needlessly penalized due to not having similar wording to a corresponding reference sentence.

There are MT evaluation metrics which utilize paraphrasing to improve the accuracy of MT evaluation ([24], [27]). Only one of them – METEOR [10] is available for the Czech language. However, its paraphrase tables are so noisy that they actually harm the performance of the metric [2], as it can award mistranslated and even untranslated words.

3 Data

We perform our experiments on data from the English-to-Czech translation task of WMT12 [8]. The data set contains 13 files with Czech outputs of MT systems and one file with corresponding reference sentences.

The human evaluation of system outputs is available as a relative ranking of performance of five systems for a sentence. We compute the absolute score of each MT system by the "> others" method [6]. It is computed as $\frac{wins}{wins+losses}$. We refer to this score as human judgment from now on.

We use two available sources of Czech paraphrases – the Czech WordNet 1.9 PDT [19] and the Czech Meteor Paraphrase Tables [9]. Czech WordNet 1.9 PDT contains high quality lemmatized paraphrases, but it is too small for our purposes.

On the other hand, the Czech Meteor Paraphrase tables are large but very noisy. For example, the following pairs are selected as paraphrases: *na poloostrově* (in a peninsula) – *šimpanzím mlékem* (milk of a chimpanzee), *gates* – *vrata* (gates) or *1873* – *pijavice* (a leech). We attempt to reduce the noise in the following way:

1. We keep only pairs consisting of single words, since we were not successful in reducing the noise effectively for the multi-word paraphrases. [3]
2. We perform morphological analysis using Morče [25] and replace the word forms with their lemmas.
3. We keep only pairs of different lemmas.
4. We dispose of pairs of words that differ in their parts of speech.
5. We dispose of pairs of words that contain an unknown word (typically a foreign word).

The last two rules have a single exception – paraphrases consisting of numeral and corresponding digits, e.g., *osmnáct* (eighteen) and *18*.[1] These paraphrases are very common in the data.

This way we reduce almost 700k pairs of paraphrases to only 32k couples of lemmas. All previous examples of incorrect paraphrases were removed. We refer to this new lemmatized paraphrase table as filtered Meteor.

4 Moses

Moses [14] is a freely available statistical machine translation engine. In a nutshell, statistical machine translation involves the following phases: creating language and translation models, parameter tuning and decoding. We use Moses in the phrase-based setting.

A language model is responsible for a correct word order and grammatical correctness of the translated sentence. A translation model (phrase table) supplies all possible translations of a word or a phrase. Models are assigned weights which are learned during the parameter tuning phase.

During the decoding phase, all these models are combined to maximize $\sum_i \lambda_i \phi_i(\bar{f}, \bar{e})$, where λ_i is a weight of a the sub-model ϕ_i and $\bar{f}, \bar{e}$ is a hypothesis and source sentence, respectively. In our case, we want to make a reference sentence closer to a corresponding machine translation output – $\bar{e}$ is the reference sentence and $\bar{f}$ is a new synthetic reference.

On its own, this setting could create paraphrases, but they would be just random paraphrases of the reference sentence – their similarity in wording to our original hypotheses would not be guaranteed. Therefore, we also add a new feature for targeted paraphrasing to Moses.

[1] *Osmnáct* has the part of speech *C*, which is designated for numerals, *18* is marked with *X* meaning it is an unknown word for the morphological analyzer.

4.1 Language model

We create the language model (LM) using the SRILM toolkit [26] on the data from the Czech part of the Czech-English parallel corpus CzEng [5].

4.2 Phrase models

Each entry in Moses phrase tables contains a phrase, its translation, several feature scores (translation probability, lexical weight etc.), and optionally also alignment within the phrase and frequencies of phrases in the training data. The phrase tables are learned automatically from large parallel data. As we do not have any large corpora of Czech-Czech parallel data, we create the following two "fake" translation models for paraphrasing from our paraphrase tables.

- **Enhanced Meteor tables**
  This table was created from the Czech Paraphrase Meteor table. It was constructed via *pivoting*. [1] The pivot method is an inexpensive way of acquiring paraphrases from large parallel corpora. It is based on the assumption that two phrases that share a meaning may have a same translation in a foreign language. [11]

  Each paraphrase pair comes with a pivoting score which we adapt as a feature in out phrase table. However, this score turns out to be even worse then random selection [3], so we do not expect it to get a high weight in tuning.

  For that reason, we add our own paraphrase scores, acquired by *distributional semantics*. Distributional semantics assumes that two phrases are semantically similar if their contextual representations are similar. [17]

  We collect all contexts (words in a window of limited size) in which Meteor paraphrases occur in the Czech National Corpus [28] and then measure context similarity (cosine distance, taking into account the number of word occurrences) for each pair of paraphrases.

  We add six scores for each pair of paraphrases according to the size of the context window used (1-3 words) and whether word order played a role in the context.

- **One-word paraphrase table**
  We first create a set of all words from Czech side of CzEng appearing at least five times to exclude rare words and possible typos. We also add all words appearing in the MT outputs and the reference sentences. Morphological analysis of the words was then performed using Morče.

  For every word x from this set, we add to this translation table every pair of words that fulfills at least on of the following requirements:

setting	reference sentence used	correlation	avg. BLEU
Baseline	original reference sentence, no paraphrasing	**0.75**	12.8
Paraphrased	paraphrased by Moses using MERT-learned weights	0.50	15.8
LM+0.2	paraphrased by Moses with LM weight increased by 0.2	0.24	9.1
LM+0.4	paraphrased by Moses with LM weight increased by 0.4	0.22	6.7

Table 1: Description of basic settings and the results - Pearson's correlation of BLEU and the human judgment, the average BLEU scores.

Source	*Paclík claims he would dare to manage the association.*
Baseline	Paclík tvrdí , že by si na vedení asociace troufl. *Paclík claims he would dare to lead the association.*
Hypothesis	Paclík tvrdí, že by se odvážil k řízení komory. *Paclík claims he would find the courage to control the chamber.*
Paraphrased	Paclík tvrdí, že by se na řízení organizace troufl. **Paclík claims he would dare to control the organization.*
LM+0.2	Paclík tvrdí, že by si troufl na řízení ekonomiky. *Paclík claims he would dare to control the economy.*
LM+0.4	Říká se, že Paclík si troufl na řídící rady. *They say that Paclík ventured to governing boards.*

Figure 1: Example of the targeted paraphrasing. The hypothesis is grammatically correct and has very similar meaning as the source sentence. The new reference is closer in wording to the hypothesis, but there is an error in a word choice. The sentences created with increased weights of the language model are both grammatically correct, but the sentence lost its original meaning.

- x,x (not every word should be paraphrased)
- x,y, if lemma of x is lemma of y (some word might have different morphology in the paraphrased sentence)
- x,y, if lemma of x and lemma of y are paraphrases according to Czech WordNet PDT 1.9.
- x,y, if lemma of x and lemma of y are paraphrases according to the filtered Meteor.

These categories constitute the first four scores in the phrase table. A pair of words gets score e if they fall in a given category, 1 (e^0) otherwise.[2] This phrase table contains more than 1,100k pairs of words.

We add another score expressing POS tag similarity between the two words. It is computed $e^{\frac{1}{a+1}}$, where a is the minimal Hamming distance between tags of the words. This probability should reflect how morphologically distant the paraphrases are.

4.3 Feature for targeted paraphrasing

In order to steer the MT decoder (translation engine) in the direction of the hypotheses, we implemented an additional feature for Moses which measures the overlap with the hypothesis. In order to keep its computation tractable during search, the overlap is defined simply as the number of words from the hypothesis confirmed by the reference translation.

Integration into the beam search algorithm used in phrase-based decoding requires us to keep track of feature state (i.e. reference words covered) to allow for correct hypothesis recombination. We also implemented an estimator of future phrase score, defined as the number of reference translation words covered by the given phrase. Our code is included in Moses.[3]

4.4 Parameter tuning

We use the minimum error rate training (MERT) [18] to find the optimal weights for our models. MERT asserts the weights to maximize the translation quality, which is measured with BLEU. We employ the reference sentences and the highest rated MT output as the parallel data for tuning.

This method, however, turned out not to be optimal for our setting. Our feature for targeted paraphrasing naturally obtains the highest weight as it provides an oracle guide towards the hypothesis.

Other important models, e.g. the language model, get comparably very small weights. The paraphrased sentences tend to be closer to the hypothesis, but not grammatically correct. Therefore, we experiment with increasing the weight of the language model manually.

[2] Phrase-table scores are considered log-probabilities.

[3] `https://github.com/moses-smt/mosesdecoder/`

setting	reference sentence used	correlation	avg. BLEU
Lexical	Only one-word paraphrase table	0.56	15.1
Lexical & LM+0.2	**Lexical** and LM weight increased by 0.2	0.33	9.5
Monotone	**Lexical** and monotone translation	**0.61**	18.1

Table 2: Additional settings and the results – Pearson's correlation and the average BLEU scores.

5 Results

We compare four different basic settings, the results are presented in Table 1 as the Pearson's correlation coefficient of BLEU and the human judgment. A visualization of the results is shown in Figure 2. The baseline score is not exceeded by any of our paraphrasing methods, in contrast to our previous results ([2], [3]).

There are several reasons for the clear decrease in correlation with paraphrased references. Hypotheses generated by the **Paraphrased** setting, while obtaining a significantly higher BLEU score, were mostly ungrammatical and reduced the correlation of our metric.

The small weight of the language model seems to be the problem, but its increase brings even more chaos. It creates hypotheses which are nice and grammatically correct but often wholly unrelated to the source sentence.

This shows that our paraphrase table noise filtering was by no means sufficient and there is still a lot of noise in our phrase tables. Furthermore, the MT output might be far from being a correct sentence – given the high weight for the targeted paraphrase feature, we essentially transform the correct reference sentences to incorrect hypotheses at all cost, using our noisy phrase tables.

Our targeting feature is also not ideal – it ignores word order and operates only on the word level (it does not model phrases). Ungrammatical translations with scrambled word order are considered perfectly fine so long as the translation contains the same words as the reference. So while the feature does provide a kind of oracle, it does not guarantee reaching the best possible translation in terms of BLEU score, let alone a grammatical translation.

Another problem is illustrated by very small weights assigned to our translation models. In fact, the highest weight was assigned to the tag similarity feature. This shows that our model features (Meteor score and distributional similarity scores) fail to distinguish good paraphrases from the noise.

The combination of noise in the translation tables and the boosted language model then caused that during the decoding phase, the most common paraphrase according to the language model with a similar tag got the preference.

Figure 1 represents an example of our paraphrasing method. The hypothesis is grammatically correct and has a very similar meaning as the reference sentence. The new paraphrased reference is slightly closer in wording to the hypothesis, but there is an error due to a bad word choice. The boosted language model reduces errors, however the meaning of the sentences is shifted. In the **LM+0.4** setting, they also differ a lot in wording from both the hypothesis and the reference sentence.

Based on such poor results, we decided to experiment with three more settings (see Table 2). We omit the Enhanced Meteor tables as they brought most of the noise to the translation. One of the common errors using the **Paraphrased** setting is scrambled word order (often, punctuation appeared in the middle of the sentences). We attempt to fix that by using monotone translation (i.e. by disabling reordering).

These constraints improve the correlation with human judgment. However, they still do not overcome the baseline results.

6 Conclusion

We experiment with paraphrasing using the phrase-based machine translation system Moses. We show that it is a universal tool that can be used for other purposes than machine translation directly. Within Moses, we introduced a new feature for targeted paraphrasing and artificial phrase tables for paraphrasing.

However, our results are inconclusive and the correlation with human judgment drops. It is caused mainly by the high amount of noise in our translation tables and not well balanced trade-off between paraphrasing and the language model.

7 Future Work

Based on our results, Moses does not seem to be the optimal tool for our task, especially unless we have at our disposal better paraphrasing tables. A new paraphrase database PPDB [12] for Czech language should be released any time now.

Furthermore, there may be a better solution than a phrase-based translation system, namely Treex [21], a highly modular NLP software system. Treex was developed for TectoMT, which is a rule-based machine translation system that operates on deep syntactic layer.

Treex implements the stratificational approach to language, adopted from the Functional Generative Description theory [23] and its later extension by the Prague Dependency Treebank [4]. It represents sentences in four layers: word layer, morphological layer, shallow-syntax layer and deep-syntax layer (tectogrammatical layer).

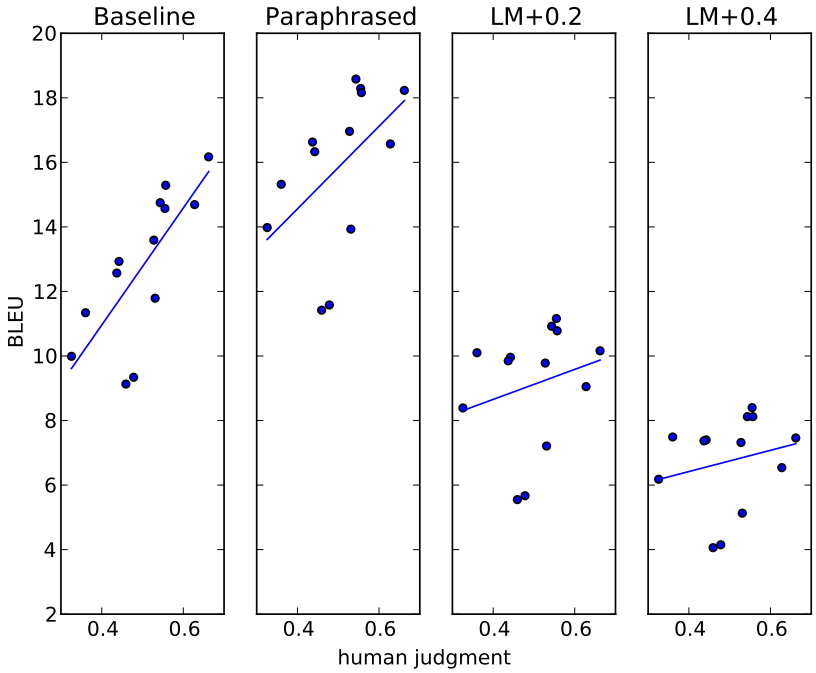

Figure 2: Visualization of BLEU and human judgment for the four basic settings. We add the linear regression lines to better demonstrate the linear correlation.

We can transfer both hypothesis and reference sentence to the morphological layer, where we can extract lemmas that appear in only one of the sentences. Those after filtering according to our paraphrase tables represent candidates for substitution. Furthermore, we are able to transfer a reference sentence to a tectogrammatical layer, where we can replace individual lemmas from the hypothesis with their paraphrases and corresponding grammatemes. Then we transfer the altered reference sentence back to the word layer.

This way should easily overcome some of the problems that appear when paraphrasing using Moses. First of all, we only compare two sentences and there is less space for the noise to interfere. Also there is highly developed machinery to avoid ungrammatical sentences. We can change only parts of sentences that are dependent on the changed word, thus keeping the rest of the sentence correct and creating more conservative reference sentences.

8 Acknowledgment

We would like to thank Ondřej Bojar for his helpful suggestions and technical advice within the NPFL101 class. This research was supported by the following grants: 1356213 of the Grant Agency of the Charles University, SVV project number 260 104 and FP7-ICT-2011-7-288487 (MosesCore). 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

[1] Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[2] Petra Barančíková. Parmesan: Meteor without Paraphrases with Paraphrased References. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT '14, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics.

[3] Petra Barančíková, Rudolf Rosa, and Aleš Tamchyna. Improving Evaluation of English-Czech MT through Paraphrasing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1711.

[4] Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague Dependency Treebank 3.0, 2013.

[5] Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The Joy

of Parallelism with CzEng 1.0. In *Proc. of LREC*, pages 3921–3928. ELRA, 2012.

[6] Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar F. Zaidan. 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, 2011. Association for Computational Linguistics.

[7] Ondřej Bojar, Kamil Kos, and David Mareček. 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, 2010. Association for Computational Linguistics.

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

[9] Michael Denkowski and Alon Lavie. 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*, 2010.

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

[11] Helge Dyvik. Translations as semantic mirrors: from parallel corpus to wordnet. In *Proceedings of the Workshop Multilinguality in the lexicon II at the 13th biennial European Conference on Artificial Intelligence (ECAI'98)*, pages 24–44, Brighton, UK, 1998.

[12] Juri Ganitkevitch and Chris Callison-Burch. The Multilingual Paraphrase Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[13] David Kauchak and Regina Barzilay. 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, 2006. Association for Computational Linguistics.

[14] 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. 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, 2007. Association for Computational Linguistics.

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

[16] Nitin Madnani. *The Circle of Meaning: From Translation to Paraphrasing and Back*. PhD thesis, Department of Computer Science, University of Maryland College Park, 2010.

[17] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[18] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

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

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. Association for Computational Linguistics.

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

[22] Rudolf Rosa, David Mareček, and Ondřej Dušek. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 362–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

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

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

[25] Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha, 2007.

[26] Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. pages 901–904, 2002.

[27] Liang Zhou, Chin yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. PARAEVAL: Using paraphrases to evaluate summaries automatically. In *IN: PROCEEDINGS OF HLT-NAACL*, pages 447–454, 2006.

[28] Ústav Českého národního korpusu FF UK. Český národní korpus - SYN2010. Praha 2010. Available at WWW: http://www.korpus.cz.