A New Syntactic Metric for Evaluation of Machine Translation

Melania Duma
Department of Computer Science
University of Hamburg
Vogt-Kölln-Straße 30
22527 Hamburg
duma@informatik.uni-hamburg.de

Cristina Vertan
Faculty for Language, Literature and Media
University of Hamburg
Von Melle Park 6
20146 Hamburg
cristina.vertan@uni-hamburg.de

Wolfgang Menzel
Department of Computer Science
University of Hamburg
Vogt-Kölln-Straße 30
22527 Hamburg
menzel@informatik.uni-hamburg.de

Abstract

Machine translation (MT) evaluation aims at measuring the quality of a candidate translation by comparing it with a reference translation. This comparison can be performed on multiple levels: lexical, syntactic or semantic. In this paper, we propose a new syntactic metric for MT evaluation based on the comparison of the dependency structures of the reference and the candidate translations. The dependency structures are obtained by means of a Weighted Constraints Dependency Grammar parser. Based on experiments performed on English to German translations, we show that the new metric correlates well with human judgments at the system level.

1 Introduction

Research in automatic machine translation (MT) evaluation has the goal of developing a set of computer-based methods that measure accurately the correctness of the output generated by a MT system. However, this task is a difficult one mainly because there is no unique reference output that can be used in the comparison with the candidate translation. One sentence can have several correct translations. Thus, it is difficult to decide if the deviation from an existing reference translation is a matter of style (the use of synonymous words, different syntax etc.) or a real translation error.

Most of the automatic evaluation metrics developed so far are focused on the idea of lexical matching between the tokens of one or more reference translations and the tokens of a candidate translation. However, structural similarity between a reference translation and a candidate one cannot be captured by lexical features. Therefore, research in MT evaluation experiences a gradual shift of focus from lexical metrics to structural ones, whether they are syntactic or semantic or a combination of both.

This paper introduces a new syntactic automatic MT evaluation method. At this stage of research the new metric is evaluating translations from any source language into German. Given that a set of constraint-based grammar rules are available for that language, extensions to other target languages are anytime possible. The chosen tool for providing syntactic information for German is the Weighted Constraints Dependency Grammar (WCDG) parser (Menzel and Schröder, 1998), which is preferred over other parsers because of its robustness to ungrammatical input, as it is typical for MT output. The rest of this paper is organized as follows. In Section 2 the state of the art in MT evaluation is presented, while in Section 3 the new syntactic metric is described. The experimental setup and results are presented in Section 4. The last section deals with the conclusions and future work.

2 State of the art

Automatic evaluation of MT systems relies on the existence of at least one reference[1] created by a human annotator. Using an automatic method of evaluation a score is computed, based on the similarity between the output of the MT system and the reference. This similarity can be computed at different levels: lexical, syntactic or semantic. At the lexical level, the metrics developed so far can be divided into two major categories: n-gram based and edit distance based.

[1] We will use the term reference for the reference translation and the term translation for the candidate translation.

Proceedings of the ACL Student Research Workshop, pages 130–135,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

Among the n-gram based metrics, one of the most popular methods of evaluation is BLEU (Papineni et al., 2001). It provides a score that is computed as the summed number of n-grams shared by the references and the output, divided by the total number of n-grams. Lexical metrics that use the edit distance are constructed using the Levenshtein distance applied at the word level. Among these metrics, WER (Niessen et al., 2000) is the one which is used more frequently; it calculates the minimal number of insertion, substitutions and deletions needed to transform the candidate translation into a reference.

Metrics based on lexical matching suffer from not being able to consider the variation encountered in natural language. Thus, they reward a low score to an otherwise fluent and syntactically correct candidate translation, if it does not share a certain number of words with the set of references. Because of this, major disagreements between the scores assigned by BLEU and human judgments have been reported in Koehn and Monz (2006) and Callison-Burch et al. (2006). Another disadvantage is that many of them cannot be applied at the segment level, which is often needed in order to better assess the quality of MT output and to determine which improvements should be made to the MT system. Because of these disadvantages there is an increasing need for other approaches to MT evaluation that go beyond the lexical level of the phrases compared.

In Liu and Gildea (2005), three syntactic evaluation metrics are presented. The first of these metrics, the Subtree Metric (SMT), is based on determining the number of subtrees that can be found in both the candidate translation and the reference phrase structure trees. The second metric, which is a kernel-based subtree metric, is defined as the maximum of the cosine measure between the MT output and the set of references. The third metric proposed computes the number of matching n-grams between the headword chains of the reference and the candidate translation dependency trees obtained using the parser described in (Collins, 1999).

The idea of syntactic similarity is further exploited in Owczarzak et al. (2007) which uses a Lexical Functional Grammar (LFG) parser. The similarity between the translation and the reference is computed using the precision and the recall of the dependencies that illustrate the pair of sentences. Furthermore, paraphrases are used in order to improve the correlation with human judgments. Another set of syntactic metrics has been introduced in Gimenez (2008); some of them are based on analyzing different types of linguistic information (i.e. part-of-speech or lemma).

3 A new syntactic automatic metric

In this section we introduce the new syntactic metric which is based on constraint dependency parsing. In the first subsection, the WCDG parser is presented, together with the advantages of using this parser over the other ones available, while the second subsection provides a detailed description of the new metric.

3.1 Weighted Constraint Dependency Grammar Parser

Our research was performed using a dependency parser. We decided on this type of parser because, as opposed to constituent parsers, it offers the possibility of better representing non-projective structures. Moreover, it has been shown in Kuebler and Prokic (2006) that, at least in the case of German, the results achieved by a dependency parser are more accurate than the ones obtained when parsing using constituent parsers, and this is because dependency parsers can handle better long distance relations and coordination.

The goal of constraint dependency grammars (CDG) is to create dependency structures that represent a given phrase (Schröder et al., 2000) on parallel levels of analysis. A relation between two words in a sentence is represented using an edge, which connects the regent and the dependent. Edges are annotated using labels in order to distinguish between different types of relations. A constraint is made up of a logical formula that describes properties of the tree. One property, for example, that is always enforced is that no word can have more than one regent on any level at a time. During the analysis, each of the constraints is applied to every edge or every pair of edges belonging to the constructed dependency parse tree. The main advantage of using constraint dependency grammars over dependency grammars based on generative rules is that they can deal better with free word order languages (Foth, 2004). Weighted Constraint Dependency Grammar (WCDG) (Menzel and Schröder, 1998) assigns different weights to the constraints of the grammar. Every constraint in WCDG is assigned a score which is a number between 0.0 and 1.0, while the general score of a parse is calculated as the product of all the scores of all the instances of constraints that have not been satisfied. Rules that have a score of 0 are called hard rules, meaning that they cannot be ignored, which is the case of the one regent only rule mentioned earlier. The advantage of using graded constraints, as opposed to crisp ones, stems from the fact that weights allow the parser to tolerate constraint violations, which, in turn, makes the parser robust against ungrammaticality. The parser was evaluated using different types of texts, and the results show that it has an accuracy between 80% and 90% in computing correct dependency attachments depending on the type of text (Foth et al., 2004a).

The benefit of using WCDG over other parsers is that it provides further information on a parse, like the general score of the parse and the constraints that are violated by the final result. This information can be further explored in order to perform an error analysis. Moreover, because of the fact that the candidate translations are sometimes not well-formed, parsing them represents a challenge. However, WCDG will always provide a final result, in the form of a dependency structure, even though it might have a low score due to the violated constraints.

3.2 Description of the metric

In order to define a new syntactic metric for MT evaluation, we have incorporated the WCDG parser in the process of evaluation. Because the output of the WCDG parser is a dependency tree, we have looked into techniques of measuring how similar two trees are. Our aim was to determine whether a tree similarity metric applied on the two dependency parse trees would prove to be an efficient way of capturing the similarity between the reference and the translation. Let us consider this example, in which the reference sentence is *“Die schwarze Katze springt schnell auf den roten Stuhl.*”(*engl. The black cat jumps quickly on the red chair*) and the candidate translation is *“Auf den roten Stuhl schnell springt die schwarze Katze”(engl. On the red chair quickly jumps the red cat)*. Even though the word order of the two segments is quite different, and the translation has an incorrect syntax, they roughly have the same meaning. We present in Figure 1 the dependency parse trees obtained using WCDG for the sentences considered. We can observe that the general structure of the translation is similar to that of the reference, the only difference being the reverse order between the left subtree and the right subtree. The tree similarity measure that we chose to use was the All Common Embedded Subtrees (ACET) (Lin et al., 2008) similarity. Given a tree *T*, an embedded subtree is obtained by removing one or more nodes, except for the root, from the tree *T*. The idea behind ACET is that, the more substructures two trees share, the more similar they are. Therefore, ACET is defined as the number of common embedded subtrees shared between two trees. The results reported in Lin et al. (2008) show that ACET outperforms tree edit distance (Zhang and Shasha, 1989) in terms of efficiency.

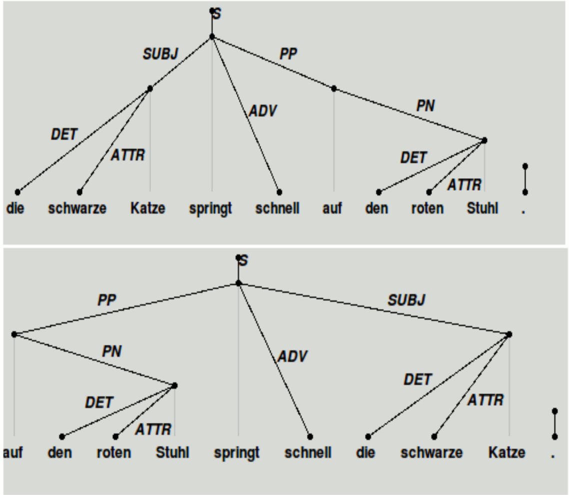

Figure 1. Example of dependency parse trees for reference and candidate translations

In our experiments, we have applied the ACET algorithm, and computed the number of common embedded subtrees between the dependency parse trees of the hypothesis and the reference. Because of the additional information provided by the parsing, pre-processing of the output of the WCDG parser was necessary in order to transform the dependency tree into a general tree. We first removed the labels assigned to every edge, but maintained the nodes and the left to right order between them.

In the following, we will refer to the new proposed metric using CESM (Common Embedded Subtree Metric). CESM was computed using the precision, the recall and the F-measure of the common embedded subtrees of the reference and the translation:

$$precision = \frac{ACET\ (tree_{ref}, tree_{hyp})}{ACET(tree_{hyp}, tree_{hyp})}$$

$$recall\ = \frac{ACET\ (tree_{ref}, tree_{hyp})}{ACET(tree_{ref}, tree_{ref})}$$

$$CESM = F_1 = \frac{2 * precision * recall}{precision + recall}$$

where $tree_{ref}$ and $tree_{hyp}$ represent the preprocessed dependency trees of the reference and the hypothesis translations.

4 Experimental setup and evaluation

In order to determine how accurate CESM is in capturing the similarity between references and translations, we evaluated it at the system level and at the segment level. The evaluation was conducted using data provided by the NAACL 2012 WMT workshop (Callison-Burch et al., 2012). The test data for the workshop consisted of 99 translated news articles in English, German, French, Spanish and Czech.

At the system level, the initial German test set provided at the workshop was filtered according to the length of segments. This was done in order to limit the time requirements of WCDG. As a result, 500 segments with a length between 50 and 80 characters were extracted from the German reference file. In the next step, we arbitrarily selected the outputs of 7 of the 15 systems that were submitted for evaluation in the English to German translation task: DFKI (Vilar, 2012), JHU (Ganitkevitch et al., 2012), KIT (Niehues et al., 2012), UK (Zeman, 2012) and three anonymized system outputs referred to as OnlineA, OnlineB, OnlineC.

After this initial step of filtering the data, the 7 systems were evaluated by calculating the CESM score for every pair of reference and translation segments corresponding to a system. The average scores obtained are depicted in Table 1. Evaluation of the metric at the system level was performed by measuring the correlation of the CESM metric with human judgments using Spearman's rank correlation coefficient ρ:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where n represents the number of MT systems considered during evaluation, and d_i^2 represents the difference between the ranks, assigned to a system, by the metric and the human judgments. The minimum value of ρ is -1, when there is no correlation between the two rankings, while the maximum value is 1, when the two rankings correlate perfectly (Callison-Burch et al., 2012). In order to compute the ρ score, the scores attributed to every system by CESM, were converted into ranks. From the different ranking strategies that were presented by the WMT12 workshop, the standard ranking order was chosen. The ρ rank correlation coefficient was calculated as being ρ = 0.92, which shows there is a strong correlation between the results of CESM and the human judgments. In order to better assess the quality of CESM, the test set was also evaluated using NIST (Doddington, 2002), which managed to obtain the same rank correlation coefficient of ρ = 0.92.

No.	System name	CESM score	NIST score
1	DFKI	0.069	4.7709
2	JHU	0.073	4.9904
3	KIT	0.090	5.1358
4	OnlineA	0.093	5.3039
5	OnlineB	0.091	5.3039
6	OnlineC	0.085	4.8022
7	UK	0.075	4.6579

Table 1. System level evaluation results

The first step in evaluating at the segment level was filtering the initial test set provided by the WMT12 workshop. For this purpose, 2500 reference and translation segments were selected with a length between 50 and 80 characters. The Kendall tau rank correlation coefficient was calculated in order to measure the correlation with human judgments, where Kendall tau (Callison-Burch et al., 2012) is defined as:

$$\tau = \frac{number\ concordant\ - number\ discordant}{number\ total\ pairs}$$

In order to compute the value of Kendall tau, we determined the number of concordant pairs and the number of discordant pairs of judgments. Similarly to the guideline followed during the WMT12 workshop (Callison-Burch et al., 2012), we penalized ties given by CESM and ignored ties assigned by the human judgments. The obtained result was a correlation of 0.058. As a term of comparison, the highest correlation for segment level reported in Callinson-Burch et al. (2012) was 0.19 obtained by TerrorCat (Fishel et al., 2012) and the lowest was BlockErrCats (Popovic, 2012) with 0.040. However, these results were obtained by evaluating on the entire test set. The rather low correlation result we obtained can be partially explained by the fact that only one judgment of a pair of reference and translation was taken into account. It will be interesting to see how the averaging of the ranks of a translation influences the correlation coefficient.

5 Conclusions and future work

In this paper, a new evaluation metric for MT was introduced, which is based on the comparison of dependency parse trees. The dependency trees were obtained using the WCDG German parser. The reason why we chose this parser was that, due to its architecture, it is able to handle ungrammatical and ambiguous input data. The experiments conducted so far show that using the data made available at the NAACL 2012 WMT workshop, CESM correlates well with the human judgments at the system level. One of the future experiments that we intend to perform is to assess metric quality on the entire evaluation set. Moreover, we plan to compare CESM with other tree-based MT metrics. Furthermore, the WMT12 workshop offers different ranking possibilities, like the ones presented in Bojar et al (2011) and in Lopez (2012). It will be determined how much are the segment level evaluation results influenced by these ranking orders.

One limitation of the proposed metric is that, at the moment it is restricted to translations from any source language to German as a target language. Because of this reason, we plan to extend the metric to other languages and see how well it performs in different settings. In further experiments we also intend to test CESM using statistical based dependency parsers, like the Malt Parser (Nivre et al., 2007) and the MST parser (McDonald et al., 2006), in order to decide whether the choice of parser influences the performance of the metric.

Another approach that we will explore for improving CESM is to compare dependency parse trees using the base form and the part-of-speech of the tokens, instead of the exact lexical match. We will try this approach in order to avoid penalizing lexical variation.

The accuracy of CESM can be further increased by the use of paraphrases, which can be obtained by using a German thesaurus or a lexical resource like GermaNet (Hamp and Feldweg, 1997). Furthermore, a technique like the one described in Owczarzak (2008) can be implemented for generating domain specific paraphrases. The results reported show that the use of this kind of paraphrases in order to produce new references has increased the BLEU score, therefore this is an approach that will be further investigated.

Acknowledgments

This work was funded by the University of Hamburg Doctoral Fellowships in accordance with the Hamburg Act for the Promotion of Young Researchers and Artists (HmbNFG), and the EAMT Project "Using Syntactic and Semantic Information in the Evaluation of Corpus-based Machine Translation".

Reference

O. Bojar, M. Ercegovčević, M Popel and O. Zaidan. 2011. *A Grain of Salt for the WMT Manual Evaluation*. Proceedings of the Sixth Workshop on Statistical Machine Translation.

C. Callison-Burch, M. Osborne and P. Koehn. 2006. *Re-evaluating the Role of Bleu in Machine Translation Research*. Proceedings of EACL-2006.

C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut and L. Specia. 2012. *Findings of the 2012 Workshop on Statistical Machine Translation*. Proceedings of WMT12.

M. J. Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

G. Doddington. 2002. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*. Proceedings of the 2nd International Conference on Human Language Technology.

K. Foth. 2004. *Writing weighted constraints for large de-pendency grammars*. Recent Advances in De-pendency Grammar, Workshop COLING 2004.

K. Foth, M. Daum and W. Menzel. 2004a. *A broad-coverage parser for German based on defeasible constraints*. KONVENS 2004, Beiträge zur 7, Konferenz zur Verarbeitung natürlicher Sprache, Wien.

K. Foth, M. Daum and W. Menzel. 2004b. *Interactive grammar development with WCDG*. Proc. of the 42nd Annual Meeting of the Association for Com-putational Linguistics.

K. Foth, T. By and W. Menzel. 2006. *Guiding a con-straint dependency parser with supertags*. Proceedings of the 21st Int. Conf. on Computational Linguistics.

M. Fishel, R. Sennrich, M. Popovic and O. Bojar. 2012. *TerrorCat: a translation error categorization-based MT quality metric*. Proceedings of the Seventh Workshop on Statistical Machine Translation.

J. Ganitkevitch, Y. Cao, J. Weese, M. Post and C. Callison-Burch. 2012. *Joshua 4.0: Packing, PRO, and paraphrases*. Proceedings of the Seventh Workshop on Statistical Machine Translation.

J. Gimenez. 2008. *Empirical Machine Translation and its Evaluation*. Ph. D. thesis.

B. Hamp and H. Feldweg. 1997. *GermaNet - a Lexical-Semantic Net for German*. Proc. of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.

P. Koehn and C. Monz. 2006. *Manual and Automatic Evaluation of Machine Translation between European Languages*. NAACL 2006 Workshop on Statistical Machine Translation.

P. Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

S. Kübler and J. Prokic. 2006. *Why is German Dependency Parsing more Reliable than Constituent Parsing?*. Proceedings of the Fifth International Work-shop on Treebanks and Linguistic Theories.

Z. Lin, H. Wang, S. McClean and C. Liu. 2008. *All Common Embedded Subtrees for Measuring Tree Similarity*. International Symposium on Computational Intelligence and Design.

D. Liu and D. Gildea. 2005. *Syntactic Features for Evaluation of Machine Translation*. ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.

A. Lopez. 2012. *Putting human assessments of machine translation systems in order*. Proceedings of the Seventh Workshop on Statistical Machine Translation.

R. McDonald, K. Lerman and F. Pereira. 2006. *Multilingual Dependency Parsing with a Two-Stage Discriminative Parser*. Tenth Conference on Computational Natural Language Learning.

W. Menzel and I. Schröder. 1998. *Decision Procedures for Dependency Parsing Using Graded Constraints*. Workshop On Processing Of Dependency-Based Grammars.

J. Niehues, Y. Zhang, M. Mediani, T. Herrmann, E. Cho and A. Waibel. 2012. *The karlsruhe institute of technology translation systems for the WMT 2012*. Proceedings of the Seventh Workshop on Statistical Machine Translation.

S. Niessen, F. J. Och, G. Leusch and H. Ney. 2000. *An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research*. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC).

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov and E. Marsi. 2007. *MaltParser: A language-independent system for data-driven dependency parsing*. Natural Language Engineering.

K. Owczarzak, J. van Genabith and A. Way. 2007. *Dependency-based automatic evaluation for machine translation*. Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation.

K. Owczarzak. 2008. *A Novel Dependency-Based Evaluation Metric for Machine Translation*, Ph.D. thesis.

K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. 2001. *Bleu: a method for automatic evaluation of machine translation*. RC22176 (Technical Report), IBM T.J. Watson Research Center.

M. Popovic. 2012. *Class error rates for evaluation of machine translation output*. Proceedings of the Seventh Workshop on Statistical Machine Translation.

I. Schröder, W. Menzel, K. Foth and M. Schulz. 2000. *Modeling dependency grammar with restricted constraints*. Traitement Automatique des Langues.

I. Schröder, H. Pop, W. Menzel and K. Foth. 2001. *Learning grammar weights using genetic algorithms*. Proceedings Euroconference Recent Advances in Natural Language Processing.

I. Schröder. 2002. *Natural Language Parsing with Graded Constraints*. Ph.D. thesis, Dept. of Computer Science, University of Hamburg.

D. Vilar. 2012. *DFKI's SMT system for WMT 2012*. Proceedings of the Seventh Workshop on Statistical Machine Translation.

D. Zeman. 2012. *Data issues of the multilingual translation matrix*. Proceedings of the Seventh Workshop on Statistical Machine Translation.

K. Zhang and D. Shasha. 1989. *Simple fast algorithms for the editing distance between trees and related problems*. SIAM Journal on Computing.