VERTa: a Linguistically-motivated Metric at the WMT15 Metrics Task

Elisabet Comelles
GRIAL
Universitat de Barcelona (UB)
Barcelona
Spain
elicomelles@ub.edu

Jordi Atserias
IXA Group
University of the Basque Country
(UPV/EHU)
Spain
jordi_atserias001@ehu.eus

Abstract

This paper describes VERTa's submission to the 2015 EMNLP Workshop on Statistical Machine Translation. VERTa is a linguistically-motivated metric that combines linguistic features at different levels. In this paper, VERTa is described briefly, as well as the three versions submitted to the workshop: VERTa-70Adeq30Flu, VERTa-EQ and VERTa-W. Finally, the experiments conducted with the WMT14 data are reported and some conclusions are drawn.

1 Introduction

In the last decade Automatic Machine Translation (MT) Evaluation has become a key field in Natural Language Processing due to the amount of texts that are translated over the world and the need for a quick, reliable and inexpensive way to evaluate the quality of the output text. Therefore, a large number of metrics have been developed, which range from very simple metrics to more complex ones. Within simple metrics there are those that do not use any type of linguistic information, such as BLEU (Papineni et al., 2002) which is one of the most well-known and widely used, since it is fast and easy to use. Other metrics though, rely on linguistic information used at lexical level such as METEOR (Denkowski and Lavie, 2014); at syntactic level, using either constituent analysis (Liu and Hildea, 2005) or dependency analysis (Owczarzack et al., 2007a and 2007b; He et al., 2010); while others use more complex information such as semantic roles (Giménez and Márquez, 2007 and 2008; Lo et al., 2012). However, all these metrics focus on partial aspects of language which might lead to a biased evaluation. As a consequence, in the last years researchers have been exploring different ways to combine a wide variety of linguistic features, either using machine-learning techniques (Leusch and Ney, 2009; Albrecht and Hwa, 2007a and 2007b; Gautam and Bhattacharyya, 2014; Joty et al., 2014) or in a more simple and straightforward way (Giménez, 2008; Giménez and Márquez, 2010; Specia and Giménez, 2010, González et al., 2014). Nevertheless, little research has been carried out in order to explore the suitability of the linguistic features used and how they should be combined, from a linguistic point of view. Therefore, this paper proposes a new version of VERTa, a linguistically-motivated metric (Comelles and Atserias, 2014) which uses a wide variety of linguistic features at different levels and which aims at moving away from a biased evaluation and providing a more holistic approach to MT evaluation. Last year VERTa participated in the WMT15 and achieved promising results at system level, this year we would like to improve the metric's performance at segment level. To this aim, a Language Model Module has been added, as well as a NERC component.

In this paper we provide a brief description of the different modules in VERTa and how they are combined, section 3 present the three versions submitted to the WMT15 and reports the experiments performed with WMT14 data into English, and finally in section 4 some conclusions are drawn.

2 VERTa: A Linguistically-motivated Metric

VERTa claims to be a linguistically-motivated metric because before its development a thorough analysis was carried out in order to identify those linguistic phenomena that an MT evalu-

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 366–372,
Lisboa, Portugal, 17-18 September 2015.

ation metric should take into account when evaluating MT output by means of reference translations. With the results of this analysis (Comelles, 2015) we decided on the linguistic features that would be more appropriate and on how they should be combined depending on whether Adequacy or Fluency was evaluated. Therefore, VERTa consists of six modules which can work independently or in combination: *Lexical Similarity Module (L), Morphological Similarity Module (M), N-gram Similarity Module (N), Dependency Similarity Module (D), Semantic Similarity Module (S)* and *Language Model (LM) Module*.

All metrics use a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triples, n-grams, etc) as shown below.

$$P = \frac{\sum_{\partial \in D} W_{\partial} * nmatch_{\partial}(\nabla(h))}{|\nabla(h)|}$$

$$R = \frac{\sum_{\partial \in D} W_{\partial} * nmatch_{\partial}(\nabla(r))}{|\nabla(r)|}$$

Where *r* is the reference, *h* is the hypothesis and ∇ is a function that given a segment will return the elements of each level (e.g. words at lexical level and triples at dependency level). *D* is the set of different functions to project the level element into the features associated to each level, such as word-form, lemma or partial-lemma at lexical level. $nmatch_{\partial}()$ is a function that returns the number of matches according to the feature ∂ (i.e. the number of lexical matches at the lexical level or the number of dependency triples that match at the dependency level). Finally, *W* is the set of weights [0 1] associated to each of the different features in a particular level in order to combine the different kinds of matches considered in that level.

Next, all modules forming VERTa are described.

2.1 Lexical Similarity Module

The Lexical Module matches lexical items in the hypothesis and reference sentences. This module does not only use superficial information such as the wordform, but it also takes into account lemmatization and lexical semantics. Hence, different types of matches are allowed and applied in the order established in Table 1. In addition, different weights can be assigned depending on their importance as regard semantics.

	Match	Examples	
		HYP	REF
1	Word-form	*east*	*east*
2	Synonym[1]	*believed*	*considered*
3	Hypernym	*barrel*	*keg*
4	Hyponym	*keg*	*barrel*
5	Lemma	*is_BE*	*are_BE*
6	Part-lemma[2]	*danger*	*dangerous*

Table 1. Lexical matches and examples

2.2 Morphological Similarity Module

This module uses the information provided by the Lexical Module in combination with Part-of-Speech (PoS) tags[3].

Similar to the Lexical Similarity Module, this module matches items in the hypothesis and reference segments and a set of weights can be assigned to each type of match (see Table 2).

	Match	Examples	
		HYP	REF
1	(Word-form, PoS)	(he, PRP)	(he, PRP)
2	(Synonym, PoS)	(VIEW, NNS)	(OPINON, NNS)
3	(Hypern., PoS)	(PUBLICATION, NN)	(MAGAZINE, NN)
4	(Hypon., PoS)	(MAGAZINE, NN)	(PUBLICATION, NN)
5	(LEMMA, PoS)	can_(CAN, MD)	Could_(CAN, MD)

Table 2. Morphological module matches

This module aims at making up for the broader coverage of the Lexical Module, thus preventing matches such as *invites* and *invite*, which although similar in meaning do not share the same morphosyntactic features.

2.3 Dependency Similarity Module

The Dependency Module makes it possible to capture similarities beyond the external structure of a sentence and uses dependency structures to link syntax and semantics. Thus, this module allows for identifying sentences with the same meaning but different syntactic constructions

[1] Information on synonyms, lemmas, hypernyms and hyponyms is obtained from WordNet 3.0.
[2] Lemmas that share the first four letters.
[3] The corpus has been PoS tagged using the Stanford Parser (de Marneffe et al. 2006).

(e.g. active – passive alternations), as well as changes in word order.

This module works at sentence level and follows the approach used by (Owczarzack et al., 2007a and 2007b) and (He et al., 2010) with some linguistic additions in order to adapt it to our metric combination. Similar to the Morphological Module, the Dependency Similarity metric also relies first on those matches established at lexical level − word-form, synonymy, hypernymy, hyponymy and lemma − in order to capture lexical variation across dependencies and avoid relying only on surface word-form. Then, by means of flat triples with the form Label(Head, Mod) obtained from the parser[4], four different types of dependency matches have been designed (see Table 3) and weights can be assigned to each type of match.

	Match Type	Match Descr.
1	Complete	Label1=Label2 Head1=Head2 Mod1=Mod2
2	Partial_no_label	Label1≠Label2 Head1=Head2 Mod1=Mod2
3	Partial_no_mod	Label1=Label2 Head1=Head2 Mod1≠Mod2
4	Partial_no_head	Label1=Label2 Head1≠Head2 Mod1=Mod2

Table 3. Dependency matches

In addition, VERTa also enables the user to assign different weights to the dependency categories according to the type of evaluation performed.

Finally, a set of language-dependent rules has been implemented in order to a) widen the range of syntactically-different but semantically-equivalent expressions, and b) restrict certain dependency relations (e.g. subject, object).

2.4 N-gram Similarity Module

This module matches chunks in the hypothesis and reference segments. N-grams can be calculated over lexical items (considering the information provided by the Lexical Module), over PoS and over the combination of lexical items and PoS. The n-gram length can go from bigrams to sentence-length grams. This module is particularly useful when evaluating Fluency because it deals with word order.

2.5 Semantic Similarity Module

The Semantic Similarity Module covers different features: Named Entities (NEs), Time Expression (TIMEX) and sentence polarity.

As regards NEs, the module uses Named Entity Recognition and classification (NERC[5]) and Named Entity Linking (NEL[6]). By means of NERC NEs of the same type are identified and matched, whereas NEL helps in matching NEs referring to the same entity regardless of their external form.

Regarding Time Expressions, the Stanford Temporal Tagger (Chang and Manning, 2012) is used to identify and match syntactically-different time expressions with the same referent.

Finally, following Wetzel and Bond (2012), who reported that negation might pose a problem to SMT systems, the metric checks and compares the polarity of the hypothesis and reference segments using the dictionary strategy described in Atserias et al. (2012).

It must be noticed, though, that in the different versions of VERTa submitted to the WMT15 only NERC is used since the rest of features did not prove to be very effective.

2.6 Language Model Module

This is a new module in VERTa, which dramatically differs from the rest of modules because the Language Model (LM) is only applied to the hypothesis sentence. By using a language model we aim at accounting for those segments that, even being syntactically different from their corresponding reference translations, are still fluent; in other words, we will be able to check the correct construction and plausibility of the hypothesis, even if it is very different or not included in any of the reference segments.

In this module we use the berkeleylm[7] implementation (Pauls and Klein, 2011), which allows for uploading LMs in different formats (e.g. arpa LM, google LM). In the experiments presented in section 3, the LM used is the NewsLM[8] re-

[4] Both hypothesis and reference strings are annotated with dependency relations by means of the Stanford parser (de Marneffe et al. 2006).

[5] In order to identify NEs we use the Supersense Tagger (Ciaramita and Altun, 2006).

[6] The NEL Module uses a graph-based NEL tool (Hachey, Radford and Curran, 2010) which links NEs in a text with those in Wikipedia pages.

[7] https://code.google.com/p/berkeleylm/

[8] http://www.quest.dcs.shef.ac.uk/quest_files/de-en/news.3gram.en.lm

leased in the WMT13 Quality Estimation Task as a baseline feature.

3 Experiments

The experiments reported in this section were carried out on the data released in WMT14, all languages into English. Language "all" includes Czech (cs), French (fr), German (de), Hindi (hi) and Russian (ru). All experiments were carried out at segment level and the evaluation sets provided by WMT organizers were used to calculate segment-level correlations.

Our goal in these experiments was two-fold: first, we wanted to test if the combination of Adequacy and Fluency features reported in Comelles (2015) was suitable for the ranking of sentences; and second, we wanted to study if the best weights for each module varied depending on the language pair.

3.1 Adequacy & Fluency Combination

This combination derives from the experiments reported in Comelles (2015), where VERTa was used to find the best combination of linguistic features in order to evaluate Adequacy and Fluency separately.

In those experiments we found out that in order to evaluate Adequacy the most effective modules were the Lexical Module, the Dependency Module, the N-gram Module and the Semantic Module (see Table 4). The strongest influence of the Lexical and the Dependency Modules is not surprising since the former accounts for lexical semantics and the latter links syntax and semantics. It must be highlighted that in the Dependency Module all types of matches were used in order to allow for matching different syntactic structures conveying the same meaning. As for the N-gram Module, n-grams were calculated over lexical items and the n-gram length was restricted to bigrams. Both N-gram and Semantic Modules showed a minor influence since the N-gram Module is more fluency-oriented and the Semantic Module focuses on very partial aspects of the evaluation.

As for the evaluation of Fluency, the ideal combination was achieved when the Dependency Module, the Language Model Module, the N-gram Module and the Morphological Module were combined (see Table 4). Some adjustments had to be performed in the Dependency and N-gram Modules. In the former only the Exact match was used so as to prevent matching constructions conveying similar meaning but which might not be completely grammatical. In the latter, n-grams were calculated over PoS and the n-grams length ranged from bigrams to sentence-length grams. The highest influence of the Dependency, N-gram and LM Modules is clear since they account for syntactic structures, morphosyntax and word order. On the other hand, the low impact of the Morphological Module is due to the fact that English does not show a rich inflectional morphology and SMT systems do not seem to have problem when dealing with it.

	Adequacy	**Fluency**
Module	**Weight**	**Weight**
Lexical	0.47	--
Morphological	--	0.04
Dependency	0.43	0.37
N-gram	0.05	0.29
Semantic	0.05	--
LM	--	0.30

Table 4. Modules combination for Adequacy and Fluency

Since our aim was finding the best way to combine Adequacy and Fluency, we performed several experiments until we found that the best correlation was obtained when the combination was Adequacy (0.70) and Fluency (0.30) (see Table 5). This indicates that semantics has a stronger influence than syntax even when dealing with ranking of segments.

Language Pair	**Correlation Coef.**
fr-en	0.406
de-en	0.323
hi-en	0.387
cz-en	0.268
ru-en	0.312
Average	0.339

Table 5. Kendall's Correlation for the Adequacy-Fluency Combination

After analyzing these results we decided to submit VERTa-70Adeq30Flu, which combined Adequacy and Fluency features with the weight combination reported above: Adequacy (0.70) and Fluency (0.30).

3.2 Language-dependent Weights

A second experiment was performed in order to study if the best weights of the modules varied depending on the language pair. To this aim, we tried all modules in VERTa with different weight combinations (see Table 6). Last year's data was

used to estimate the best weights for VERTa's modules by systematically testing all the different weight combinations (all integer weight combinations totaling 100 using a step of 5).

According to the results obtained, the module that influences the most in almost all language pairs (i.e. de-en, cz-en and ru-en) is the Dependency Module. This might be due to the fact that the dependency relations are a halfway stage between syntax and semantics. They help to link the surface structure of a sentence with its deep structure, closer to semantics. In addition, the Dependency Module relies on information provided by the Lexical Module which is related to lexical semantics, again escaping the word-form and moving towards meaning. The exceptions to the remarkable influence of the Dependency Module are the fr-en pair, where the LM Module shows a stronger influence than the rest of modules, and the hi-en pair, where the Lexical Module is assigned the highest weight. As for the Lexical Module, its influence is rather low for most of the languages – with the exception of the hi-en pair – however, it shows a good performance when the average correlation is calculated. Regarding the N-gram Module, its influence is similar in most language pairs (i.e. hi-en, cz-en and ru-en), as well as the average score, which might be explained by the importance of word order. The Morphological Module does not seem to be very suitable because it only proves efficient for the de-en pair, and up to a certain point for the cz-en pair. Finally, the Semantic Module does not show any impact, which might be due to the fact that only NEs were used and, as already mentioned, they only account for a very partial aspect of the translation.

Lang.	**Weight Combination**[9]						**Corr.**
	L	*M*	*D*	*N*	*S*	*LM*	
fr-en	0	10	10	10	0	70	0.427
de-en	10	20	50	10	0	10	0.323
hi-en	40	0	20	20	0	20	0.390
cz-en	10	10	50	20	10	0	0.269
ru-en	20	0	30	30	0	20	0.318
Aver.	30	0	40	20	0	10	0.339

Table 6. Kendall's Correlation for language-dependent weight combinations

[9] Weights corresponding to: Lexical Module (L), Morphological Module (M), Dependency Module (D), N-gram Module (N), Semantic Module (S) and LM Module (LM).

Given the results obtained, we decided to submit two more versions of VERTa:

- VERTa-W. This version uses the following settings, except for the fr-en pair: Lexical Module (0.30), Dependency Module (0.40), N-gram Module (0.20) and Language Model Module (0.10). The reason why these modules and weights are chosen is that they were the settings that obtained the best average correlation at segment level (see Table 6). As regards the fr-en language pair, since it showed a completely different behaviour to the rest of language pairs, different modules and weights were used. Hence the settings used for the fr-en pair are those reported in Table 6, which involve a really strong influence of the Language Model Module. Using these settings to evaluate the rest of language pairs drops the average correlation of all languages significantly, from 0.339 to 0.310.

- VERTa-EQ. In line with last year's submission, this submission combines all modules in VERTa with equal weights assigned to each module, thus combining linguistic features in a more simple and straightforward way.

3.3 Comparing Different Versions of VERTa

In this section the different versions of VERTa submitted to the WMT15 are compared to those submitted to the WMT14 (see Table 7). In addition, the best and worst systems of the 2014 edition are also included for the sake of comparison.

WMT15 results show that both VERTa-W and VERTa-70Adeq30Flu achieve similar results in the average correlation and for the hi-en language pair. However, VERTa-W performs better for the fr-en and, especially, for the ru-en pair. The reason why VERTa-W performs better for the fr-en pair is that, as explained in section 3.2, the settings used differ completely from those used for the rest of language pairs, since experiments showed that a higher influence of the LM Modules was advisable. As for the ru-en pair, the more efficient performance might be due to the fact that in VERTa-W the Morphological Module and the Semantic Module are disregarded, which coincides with the best setting for ru-en shown in Table 6.

On the other hand, VERTa-70Adeq30Flu performs better for the de-en and cz-en pairs. In both cases this is due to the fact that both

Metric		fr-en	de-en	hi-en	cz-en	ru-en	Average
WMT14	**VERTa-W**	0.399	0.321	0.386	0.263	0.315	0.337
	VERTa-EQ	0.407	0.315	0.384	0.263	0.312	0.336
	Best-WMT14	0.433	0.380	0.434	0.328	0.355	0.386
	Worst-WMT14	0.005	0.001	0.000	0.002	0.001	0.002
WMT15	**VERTa-W**	**0.408**	0.321	**0.387**	0.262	**0.316**	**0.339**
	VERTa-EQ	0.393	0.313	0.370	0.260	0.292	0.325
	VERTa-70Adeq30Flu	0.406	**0.323**	**0.387**	**0.268**	0.312	**0.339**

Table 7. Comparison between VERTa's submission to WMT14 and WMT15

Morphological and Semantic Modules are used in this version which, according to the weight combination in Table 6, allows for a better performance of the metric when evaluating those two language pairs.

As for VERTa-EQ, the last version submitted to WMT15, its performance is the lowest of the three submissions. This is a direct consequence of assigning the same weights to all modules, when experiments have clearly shown that there are some modules more effective than others.

As regards the difference between WMT14 and WMT15 submissions, unfortunately our results have not improved as much as we expected. Nevertheless, both VERTa-W and VERTa-70Adeq30Flu improve their average score in 0.002, from 0.337 to 0.339. As for the scores obtained for each language pair, the cz-en pair undergoes the most remarkable improvement, moving from 0.263 up to 0.268.

4 Conclusions and Future Work

In this paper we have described VERTa, a linguistically-motivated MT metric and the three versions submitted to the WMT15: VERTa-70Adeq30Flu, VERTa-W and VERTa-EQ. VERTa-70Adeq30Flu combines Adequacy features and Fluency features to rank MT segments; VERTa-W uses some of the modules in VERTa with different weights assigned to each module; and finally, VERTa-EQ uses all modules in VERTa with equal weights assigned.

Two first versions of VERTa were submitted last year; however, our current submissions to WMT15 include two more modules: the first new module uses a NERC component whereas the second uses a Language Model.

By means of our experiments we have been able to study two key areas in automatic MT evaluation: a) how Adequacy and Fluency features can be used and adapted to ranking-based evaluation; and b) how VERTa behaves when different pairs of languages are considered.

Our experiments have shown that VERTa shows a stable performance for almost all language pairs evaluated, with the exception of the fr-en pair, for which the LM Module seemed to be the most effective one. Such high influence might indicate that when translating from French into English word order plays an important role and MT evaluation metrics should handle it effectively.

Finally, we have compared our new versions to the versions submitted last year, and although results are not outstanding, VERTa's performance at segment level has improved slightly, especially in the case of VERTa-70Adeq30Flu and VERTa-W.

In the future we would like to apply machine-learning techniques to the combination of modules since we think our metric could greatly benefit from this approach. In addition, since our metric uses a wide range of NLP tools, we would like to explore how NLP tool errors influence the performance of the metric.

References

J. S. Albrecth and R. Hwa. 2007. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *The Proceedings of the 45th Annual Meeting of the ACL*, Prague, Czech Republic.

J. S. Albrecth and R. Hwa. 2007. Regression for Sentence-Level MT Evaluation with Pseudo References. In *The Proceedings of the 45th Annual Meeting of the ACL*, Prague, Czech Republic.

J. Atserias, R. Blanco, J. M. Chenlo and C. Rodriquez. 2012. FBM-Yahoo at RepLab 2012, *CLEF (Online Working Notes/Labs/Workshop) 2012*, September 20, 2012.

A. X. Chang and Ch. D. Manning. 2012. SUTIME: A Library for Recognizing and Normalizing Time Expressions. *8th International Conference on Language Resources and Evaluation (LREC 2012)*.

M. Ciaramita and Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction

with a supersense sequence tagger. Empirical Methods in *Natural Language Processing (EMNLP)*.

E. Comelles. 2015. *Automatic Machine Translation Evaluation: A Qualitative Approach*. Doctoral Dissertation. University of Barcelona.

E. Comelles and J. Atserias. 2014. VERTa participation in the WMT14 Metrics Task in *Proceedings of the Ninth Workshop on Statistical Machine Translation (ACL-2014)*. Baltimore, Maryland, USA.

M.C. de Marneffe, B. MacCartney and Ch. D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses in *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation (LREC-2006)*. Genoa, Italy.

M. J. Denkowski and A. Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language in *Proceedings of the Ninth Workshop on Statistical Machine Translation (ACL-2014)*. Baltimore, Maryland, USA.

S. Gautam, and P. 2014. LAYERED: Metric for Machine Translation Evaluation in *Proceedings of the Ninth Workshop on Statistical Machine Translation (ACL-2014)*. Baltimore, Maryland, USA.

J. Giménez and Ll. Màrquez. 2007. Linguistic features for automatic evaluation of heterogeneous MT systems in *Proceedings of the 2nd Workshop on Statistical Machine Translation (ACL)*, Prague, Czech Repubilc.

J. Giménez and Ll. Màrquez. 2008. A smorgasbord of features for automatic MT evaluation in *Proceedings of the 3rd Workshop on Statistical Machine Translation (ACL)*. Columbus. OH.

J. Gimenez. 2008. *Empirical Machine Translation and its Evaluation*. Doctoral Dissertation. UPC.

J. Giménez and Ll. Márquez. 2010. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3–4),77–86. Springer.

M. González, A. Barrón-Cedeño and Ll. Márquez. 2014. IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation in *Proceedings of the Ninth Workshop on Statistical Machine Translation (ACL-2014)*. Baltimore, Maryland, USA.

B. Hachey, W. Radford and J. R. Curran. 2011. Graph-based named entity linking with Wikipedia in *Proceedings of the 12th International conference on Web information system engineering*, pages 213-226, Springer-Verlag, Berlin, Heidelberg.

Y. He, J. Du, A. Way and J. van Genabith. 2010. The DCU Dependency-based Metric in WMT-Metrics MATR 2010. *In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, Uppsala, Sweden.

S. Joty, F. Guzmán, Ll. Márquez and P. Nakov. 2014. DiscoTK: Using Discourse Structure for Machine Translation Evaluation in *Proceedings of the Ninth Workshop on Statistical Machine Translation (ACL-2014)*. Baltimore, Maryland, USA.

G. Leusch and H. Ney. 2008. BLEUSP, INVWER, CDER: Three improved MT evaluation measures. In *NIST Metrics for Machine Translation 2008 Evaluation (MericsMATR08)*, Waikiki, Honolulu, Hawaii, October 2008.

D. Liu and D. Hildea. 2005. Syntactic Features for Evaluation of Machine Translation in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor

Ch. Lo, A. K. Tumurulu and D. Wu. 2012. Fully Automatic Semantic MT Evaluation. *Proceedings of the 7th Wrokshop on Statistical Machine Translation*, Montréal, Canada, June 7-8.

K. Owczarzak, J. van Genabith and A. Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation in *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure I Statistical Translation*, Rochester, New York.

K. Owczarzak, J. van Genabith and A. Way. 2007. Labelled Dependencies in Machine Translation Evaluation in *Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic.

K. Papineni, S. Roukos, T. Ward and W. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL02). Philadelphia. PA.

A. Pauls and D. Klein. 2011. Faster and smaller *N*-gram language models in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (HLT '11), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA.

L. Specia and J. Giménez. 2010. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. *The Ninth Conference of the Association for Machine Translation in the Americas* (AMTA 2010), Denver, Colorado.

D. Wetzel and F. Bond. 2012. Enriching Parallel Corpora for Statistical Machine Translation with Semantic Negation Rephrasing. *Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Jeju, Republic of Korea.