VERTa participation in the WMT14 Metrics Task

Elisabet Comelles
Universitat de Barcelona
Barcelona, Spain
elicomelles@ub.edu

Jordi Atserias
Yahoo! Labs
Barcelona, Spain
jordi@yahoo-inc.com

Abstract

In this paper we present VERTa, a linguistically-motivated metric that combines linguistic features at different levels. We provide the linguistic motivation on which the metric is based, as well as describe the different modules in VERTa and how they are combined. Finally, we describe the two versions of VERTa, VERTa-EQ and VERTa-W, sent to WMT14 and report results obtained in the experiments conducted with the WMT12 and WMT13 data into English.

1 Introduction

In the Machine Translation (MT) process, the evaluation of MT systems plays a key role both in their development and improvement. From the MT metrics that have been developed during the last decades, BLEU (Papineni et al., 2002) is one of the most well-known and widely used, since it is fast and easy to use. Nonetheless, researchers such as (Callison-Burch et al., 2006) and (Lavie and Dekowski, 2009) have claimed its weaknesses regarding translation quality and its tendency to favour statistically-based MT systems. As a consequence, other more complex metrics that use linguistic information have been developed. Some use linguistic information at lexical level, such as METEOR (Denkowski and Lavie, 2011); others rely on syntactic information, either using constituent (Liu and Hildea, 2005) or dependency analysis (Owczarzack et al., 2007a and 2007b; He et al., 2010); others use more complex information such as semantic roles (Giménez and Márquez, 2007 and 2008a; Lo et al., 2012). All these metrics focus on partial aspects of language; however, other researchers have tried to combine information at different linguistic levels in order to follow a more holistic approach. Some of these metrics follow a machine-learning approach (Leusch and Ney, 2009; Albrecht and Hwa, 2007a and 2007b), others combine a wide variety of metrics in a simple and straightforward way (Giménez, 2008b; Giménez and Márquez, 2010; Specia and Giménez, 2010). However, very little research has been performed on the impact of the linguistic features used and how to combine this information from a linguistic point of view. Hence, our proposal is a linguistically-based metric, VERTa (Comelles et al., 2012), which uses a wide variety of linguistic features at different levels, and aims at combining them in order to provide a wider and more accurate coverage than those metrics working at a specific linguistic level. In this paper we provide a description of the linguistic information used in VERTa, the different modules that form VERTa and how they are combined according to the language evaluated and the type of evaluation performed. Moreover, the two versions of VERTa participating in WMT14, VERTa-EQ and VERTa-W are described. Finally, for the sake of comparison, we use the data available in WMT12 and WMT13 to compare both versions to the metrics participating in those shared tasks.

2 Linguistic Motivation

Before developing VERTa, we analysed those linguistic phenomena that an MT metric should cover. From this analysis, we decided to organise the information into the following groups:

- **Lexical information**. The use of lexical semantics plays a key role when comparing a hypothesis and reference segment, since it allows for identifying relations of synonymy, hypernymy and hyponymy.
- **Morphological information**. This type of information is crucial when dealing with languages with a rich inflectional morphology, such as Spanish, French or Cata-

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 368–375,
Baltimore, Maryland USA, June 26–27, 2014.

lan because it helps in covering phenomena related to tense, mood, gender, number, aspect or case. In addition, morphology in combination with syntax (morpho-syntax) is also important to identify agreement (i.e. subject-verb agreement). This type of information should be taken into account when evaluating the fluency of a segment.

- **Syntactic information.** This type of information covers syntactic structure, syntactic relations and word order.
- **Semantic information.** Named Entities (NEs), sentence polarity and time expressions are included here.

All this information described above should be taken into account when developing a metric that aims at covering linguistic phenomena at different levels and evaluate both adequacy and fluency.

3 Metric Description

In order to cover the above linguistic features, VERTa is organised into different modules: *Lexical similarity module*, *Morphological similarity module*, *Dependency similarity module* and *Semantic similarity module*. Likewise, an *Ngram similarity module* has also been added in order to account for similarity between chunks in the hypothesis and reference segments. Each metric works first individually and the final score is the Fmean of the weighted combination of the Precision and Recall of each metric in order to get the results which best correlate with human assessment. This way, the different modules can be weighted depending on their importance regarding the type of evaluation (fluency or adequacy) and language evaluated. In addition, the modular design of this metric makes it suitable for all languages. Even those languages that do not have a wide range of NLP tools available could be evaluated, since each module can be used isolated or in combination.

All metrics use a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triples, ngrams, 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.

All modules forming VERTa and the linguistic features used are described in detail in the following subsections.

3.1 Lexical module

Inspired by METEOR, the lexical module matches lexical items in the hypothesis segment to those in the reference segment taking into account several linguistic features. However, while METEOR uses word-form, synonymy, stemming and paraphrasing, VERTa relies on word-form, synonymy[1], lemma, partial lemma[2], hypernyms and hyponyms. In addition, a set of weights is assigned to each type of match depending on their importance as regards semantics (see Table 1).

	W	Match	Examples	
			HYP	REF
1	1	Word-form	*east*	*east*
2	1	Synonym	*believed*	*considered*
3	1	Hypernym	*barrel*	*keg*
4	1	Hyponym	*keg*	*barrel*
5	.8	Lemma	*is_BE*	*are_BE*
6	.6	Part-lemma	*danger*	*dangerous*

Table 1. Lexical matches and examples.

3.2 Morphological similarity module

The morphological similarity module is based on the matches established in the lexical module (except for the partial-lemma match) in combination with Part-of-Speech (PoS) tags from the annotated corpus[3]. The aim of this module is to

[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).

compensate the broader coverage of the lexical module, preventing matches such as *invites* and *invite*, which although similar in terms of meaning, do not coincide as for their morphological information. Therefore, this module turns more appropriate to assess the fluency of a segment rather than its adequacy. In addition, this module will be particularly useful when evaluating languages with a richer inflectional morphology (i.e. Romance languages).

In line with the lexical similarity metric, the morphological similarity metric establishes matches between items in the hypothesis and the reference sentence and a set of weights (W) is applied. However, instead of comparing single lexical items as in the previous module, in this module we compare pairs of features in the order established in Table 2.

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

Table 2. Morphological module matches.

3.3 Dependency similarity module

The dependency similarity metric helps in capturing similarities between semantically comparable expressions that show a different syntactic structure (see Example 1), as well as changes in word order (see Example 2).

Example 1:

HYP: *...the interior minister...*

REF: *...the minister of interior...*

In example 1 both hypothesis and reference chunks convey the same meaning but their syntactic constructions are different.

Example 2:

HYP: ***After a meeting Monday night with the head of Egyptian intelligence chief Omar Suleiman*** *Haniya said....*

REF: *Haniya said,* ***after a meeting on Monday evening with the head of Egyptian Intelligence General Omar Suleiman****...*

In example 2, the adjunct realised by the PP *After a meeting Monday night with the head of Egyptian intelligence chief Omar Suleiman* occupies different positions in the hypothesis and reference strings. In the hypothesis it is located at the beginning of the sentence, preceding the subject *Haniya*, whereas in the reference, it is placed after the verb. By means of dependencies, we can state that although located differently inside the sentence, both subject and adjunct depend on the verb.

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 have been assigned to each type of match.

W	Match Type	Match Descr.
1	Complete	Label1=Label2 Head1=Head2 Mod1=Mod2
1	Partial_no_label	Label1≠Label2 Head1=Head2 Mod1=Mod2
.9	Partial_no_mod	Label1=Label2 Head1=Head2 Mod1≠Mod2
.7	Partial_no_head	Label1=Label2 Head1≠Head2 Mod1=Mod2

Table 3. Dependency matches.

In addition, dependency categories also receive a different weight depending on how informative they are: *dep*, *det* and _[5] which receive 0.5, whereas the rest of categories are assigned the maximum weight (1).

Finally, a set of language-dependent rules has been added with two goals: 1) capturing similarities between different syntactic structures con-

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

[5] _ stands for no_dep_label

veying the same meaning; and 2) restricting certain dependency relations (i.e. subject word order when translating from Arabic to English).

3.4 Ngram similarity module

The ngram similarity metric matches chunks in the hypothesis and reference segments and relies on the matches set by the lexical similarity metric, which allows us to work not only with word-forms but also with synonyms, lemmas, partial-lemmas, hypernyms and hyponyms as shown in Example 3, where the chunks [*the situation in the area*] and [*the situation in the region*] do match, even though *area* and *region* do not share the same word-form but a relation of synonymy.

Example 3:
HYP: … *the situation in the* ***area***…
REF: … *the situation in the* ***region***…

3.5 Semantics similarity module

As confirmed by the lexical module, semantics plays an important role in the evaluation of adequacy. This has also been claimed by (Lo and Wu, 2010) who report that their metric based on semantic roles outperforms other well-known metrics when adequacy is assessed. With this aim in mind the semantic similarity module uses other semantic features at sentence level: NEs, time expressions and polarity.

Regarding NEs, we use Named-Entity recognition (NER) and Named-Entity linking (NEL). Following previous NE-based metrics (Reeder et al., 2011 and Giménez, 2008) the NER metric[6] aims at capturing similarities between NEs in the hypothesis and reference segments. On the other hand NEL[7] focuses only on those NEs that appear on Wikipedia, which allows for linking NEs regardless of their external form. Thus, *EU* and *European Union* will be captured as the same NE, since both of them are considered as the same organisation in Wikipedia.

As regards time expressions, the TIMEX metric matches temporal expressions in the hypothesis and reference segments regardless of their form. The tool used is the Stanford Temporal Tagger (Chang and Manning, 2012) which recognizes not only points in time but also duration. By means of this metric, different syntactic structures conveying the same time expression can be matched, such as *on February 3rd* and *on the third of February*.

Finally, it has been reported that negation might pose a problem to SMT systems (Wetzel and Bond, 2012). In order to answer such need, a module that checks the polarity of the sentence has been added using the dictionary strategy described (Atserias et al., 2012):

- Adding 0.5 for each weak positive word.
- Adding 1.0 for each strong positive word.
- Subtracting 0.5 for each weak negative word.
- Subtracting 1.0 for each strong negative word.

For each query term score, the value is propagated to the query term positions by reducing its strength in a factor of 1/n, where n is the distance between the query term and the polar term.

According to the experiments performed, this module shows a low correlation with human judgements on adequacy, since only partial aspects of translation are considered, whereas human judges assess whole segments. However, regardless of how well/bad the module correlates with human judgements, it proves useful to check partial aspects of the segments translated, such as the correct translation of NEs or the correct translation of negation.

3.6 Metrics combination

The modular design of VERTa allows for providing different weights to each module depending on the type of evaluation and the language evaluated. Thus following linguistic criteria when evaluating adequacy, those modules which must play a key role are the lexical and dependency module, since they are more related to semantics; whereas, when evaluating fluency those related to morphology, morphosyntax and constituent word order will be the most important. Moreover, metrics can also be combined depending on the type of language evaluated. If a language with a rich inflectional morphology is assessed, the morphology module should be given a higher weight; whereas if the language evaluated does not show such a rich inflectional morphology, the weight of the morphology module should be lower.

4 Experiments and results

Experiments were carried out on WMT data, specifically on WMT12 and WMT13 data, all languages into English. Languages "all" include French, German, Spanish and Czech for WMT12

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

[7] 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.

and French, German, Spanish, Czech and Russian for WMT13. Both segment and system level evaluations were performed. Evaluation sets provided by WMT organizers were used to calculate both segment and system level correlations.

Since VERTa has been mainly designed to assess either adequacy or fluency separately, our goal for WMT14 was to find the best combination in order to evaluate whole translation quality. Firstly we decided to explore the influence of each module separately. To this aim, all modules described above, except for the semantics one were used and tested separately. Secondly, all modules were assigned the same weight and tested in combination (VERTa-EQ). The reason why the semantics module was disregarded is that it does not usually correlate well with human judgements, as stated above. Each module was set as follows:

- Lexical module. As described above, except for the use of hypernyms/hyponyms matches that were disregarded.
- Morphological module. As described above, except for the lemma-PoS match and the hypernyms/hyponyms-PoS match.
- Dependency module. As described above.
- Ngram module. As described above, using a 2-gram length.

Finally, we used the module combination aimed at evaluating adequacy, which is mainly based on the dependency and lexical modules, but with a stronger influence of the ngram module in order to control word order (VERTa-W). Weights were manually assigned, based on results obtained in previous experiments conducted for adequacy and fluency (Comelles et al., 2012), as follows:

- Lexical module: 0.41
- Morphological module: 0
- Dependency module: 0.40
- Ngram module: 0.19

Experiments aimed at evaluating the influence of each module (see Table 4 and Table 5) show that the dependency module, in the case of WMT12 data, and the lexical module in the case of WMT13 data, are the most effective ones. However, the influence of the ngram module and the morphological module varies depending on the source language. The fact that the dependency module correlates better with human judgements than others might be due to its flexibility to capture different syntactic constructions that convey the same meaning. In addition, the good performance of the lexical module is due to the use of lexical semantic relations. On the other hand, in general the morphological module shows a better performance than the ngram one, which might be due to the type of source languages and the possible translation mistakes. All source languages are highly-inflected languages and this might cause problems when translating into English, since its inflectional morphology is not as rich as theirs. As for the low performance of the ngram module in the cs-en (especially, in WMT12 data), it might be due to the fact that Czech word order is unrestricted, whereas English shows a stricter word order and this might cause translation issues. A longer ngram distance might have been more appropriate to control word order in this case.

Module	**fr-en**	**de-en**	**es-en**	**cs-en**
Lexical	.16	.20	.18	.14
Morph.	.17	.19	.18	.12
Depend.	**.18**	**.24**	**.20**	**.17**
Ngram	.16	.17	.15	.08

Table 4. Segment-level Kendall's tau correlation per module with WMT12 data.

Module	**fr-en**	**de-en**	**es-en**	**cs-en**	**ru-en**
Lexical	**.239**	**.254**	**.294**	**.227**	**.220**
Morph.	.236	.243	.295	.214	.191
Depend.	.232	.247	.275	.220	.199
Ngram	.237	.245	.283	.213	.189

Table 5. Segment-level Kendall's tau correlation per module with WMT13 data.

Finally, two versions of VERTa were compared: the unweighted combination (VERTa-EQ) and the weighted one (VERTa-W). These two versions were also compared to some of the best performing metrics in WMT12 (see Table 6 and Table 7) and WMT13 (see Table 8 and Table 9): Spede07-pP, METEOR, SEMPOR and AMBER (Callison-Burch et al., 2012); SIMPBLEU-RECALL, METEOR and DEPREF-ALIGN[8]). As regards WMT12 data at segment level, the unweighted version achieves similar results to those obtained by the best performing metrics. On the other hand, VERTa-W's results are slightly worse, especially for fr-en and es-en pairs, which is due to the fact that the morphological module has been disregarded in this ver-

[8] http://www.statmt.org/wmt13/papers.html

sion. Regarding system level correlation, neither VERTa-EQ nor VERTa-W achieves a high correlation with human judgements.

Metric	fr-en	de-en	es-en	cs-en
Spede07-pP	**.26**	**.28**	**.26**	**.21**
METEOR	.25	.27	.24	.21
VERTa-EQ	**.26**	**.28**	**.26**	.20
VERTa-W	.24	.28	.25	.20

Table 6. Segment-level Kendall's tau correlation WMT12.

Metric	fr-en	de-en	es-en	cs-en
SEMPOR	.80	**.92**	.94	**.94**
AMBER	**.85**	.79	**.97**	.83
VERTa-EQ	.83	.71	.89	.66
VERTa-W	.79	.73	.91	.66

Table 7. System-level Spearman's rho correlation WMT12.

As for segment level WMT13 results (see Table 8), although both VERTa-EQ and VERTa-W's performance is worse than that of the two best-performing metrics, both versions achieve a third and fourth position for all language pairs, except for fr-en. As regards system level correlations (see Table 9), both versions of VERTa show the best performance for de-en and ru-en pairs, as well as for the average score.

5 Conclusions and Future Work

In this paper we have presented VERTa, a linguistically-based MT metric. VERTa allows for modular combination depending on the language and type of evaluation conducted. Although VERTa has been designed to evaluate adequacy and fluency separately, in order to evaluate whole MT quality, a couple of versions have been used: VERTa-EQ, an unweighted version that uses all modules, and VERTa-W a weighted version that uses the lexical, dependency and ngram modules.

Experiments have shown that the modules that best correlate with human judgements are the dependency and lexical modules. In addition, both VERTa-EQ and VERTa-W have been compared to the best performing metrics in WMT12 and WMT13 shared tasks. VERTa-EQ has proved to be in line with results obtained by Spede07-pP and METEOR in WMT12 at segment level, while in WMT13, both VERTa and VERTa-W occupy the third and fourth position after METEOR and DEPREF-ALIGN as regards segment level and the first position at system level.

In the future, we plan to continue working on the improvement of VERTa and use automatic tuning of module's weight in order to achieve the final version that best correlates with human judgements on ranking. Likewise, we would like to explore the use of VERTa to evaluate other languages but English and how NLP tool errors may influence the performance of the metric.

6 Acknowledgements

We would like to acknowledge Victoria Arranz and Irene Castellón for their valuable comments and sharing their knowledge.

This work has been partially funded by the Spanish Government (projects SKATeR, TIN2012-38584-C06-06 and Holopedia, TIN2010-21128-C02-02).

Metric	fr-en	de-en	es-en	cs-en	ru-en	Average
SIMPBLEU-RECALL	**.303**	**.318**	**.388**	**.260**	**.234**	**.301**
METEOR	.264	.293	.324	.265	.239	.277
VERTa-EQ	.252	.280	.318	.239	.215	.261
VERTa-W	.253	.278	.314	.238	.222	.261
DEPREF-ALIGN	.257	.267	.312	.228	.200	.253

Table 8. Segment-level Kendall's tau correlation WMT13.

Metric	fr-en	de-en	es-en	cs-en	ru-en	Average
METEOR	.984	.961	**.979**	**.964**	.789	.935
DEPREF-ALIGN	**.995**	.966	.965	**.964**	.768	.931
VERTa-EQ	.989	**.970**	.972	.936	**.814**	**.936**
VERTa-W	.989	**.980**	.972	.945	**.868**	**.951**

Table 9. System-level Spearman's rho correlation WMT13.

Reference

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.

C. Callison-Burch, M. Osborne and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the EACL 2006*.

C. Callison-Burch, P. Kohen, Ch. Monz, M. Post, R. Soricut and L. Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Montréal. Canada.

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, J. Atserias, V. Arranz and I. Castellón. 2012. VERTa: Linguistic features in MT evaluation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

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. 2011. METEOR 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems in *Proceedings of the 6th Workshop on Statistical Machine Translation (ACL-2011)*. Edinburgh, Scotland, UK.

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.

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.

A. Lavie and M. J. Denkowski. 2009. The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation, 23*.

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 and D. Wu. 2010. Semantic vs. Syntactic vs. Ngram Structure for Machine Translation Evaluation. In *Proceedings of the 4th Workshop on Syntax Semantics and Structure in Statistical Translation*. Beijing. China.

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.

F. Reeder, K. Miller, J. Doyon and J. White. 2001. The Naming of Things and the Confusion of Tongues: an MT Metric. *Proceedings of the Workshop on MT Evaluation "Who did what to whom?" at Machine Translation Summit VIII*.

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.