VERTa: Linguistic Features in MT Evaluation

Elisabet Comelles§, Jordi Atserias‡, Victoria Arranz*, Irene Castellón§
§University of Barcelona
Gran Via de les Corts Catalanes, 585
08007 Barcelona, Spain
‡Fundació Barcelona Media
Av. Diagonal, 177
08018 Barcelona, Spain
*ELDA/ELRA
55-57 rue Brillat Savarin
75013 Paris, France

E-mail: elicomelles@ub.edu, jordi.atserias@barcelonamedia.org, arranz@elda.org, icastellon@ub.edu

Abstract

In the last decades, a wide range of automatic metrics that use linguistic knowledge has been developed. Some of them are based on lexical information, such as METEOR; others rely on the use of syntax, either using constituent or dependency analysis; and others use semantic information, such as Named Entities and semantic roles. All these metrics work at a specific linguistic level, but some researchers have tried to combine linguistic information, either by combining several metrics following a machine-learning approach or focusing on the combination of a wide variety of metrics in a simple and straightforward way. However, little research has been conducted on how to combine linguistic features from a linguistic point of view. In this paper we present VERTa, a metric which aims at using and combining a wide variety of linguistic features at lexical, morphological, syntactic and semantic level. We provide a description of the metric and report some preliminary experiments which will help us to discuss the use and combination of certain linguistic features in order to improve the metric performance.

Keywords: MT evaluation, automatic metric, linguistically-based metric

1. Introduction

In the MT field, the evaluation of MT systems plays an important role both in their development and improvement. However, human evaluation is expensive and complex, and consequently, in the last decades, a wide range of automatic metrics have been developed, from which the most well-known and widely-spread is the string-based IBM's BLEU (Papineni et al 2002). However, researchers such as Callison-Burch et al. (2006) and Lavie and Dekowski (2009) have been critical to its performance and have highlighted its weaknesses in relation to translation quality and its tendency to favour statistically-based MT systems. In response to these weaknesses, more sophisticated metrics, most of them linguisticaly-motivated, have arisen. Some of them are based on lexical information, such as METEOR (Banerjee and Lavie, 2011); others rely on the use of syntax, either using constituent (Liu and Hildea 2005) or dependency analysis (Owczarzack et al. 2007a and 2007b; He et al. 2010); and others use semantic information, such as Named Entities and semantic roles (Giménez and Márquez 2007 and 2008a). All these metrics work at a specific linguistic level, but other researchers have tried to combine linguistic information, either by combining several metrics following a machine-learning approach, such as Leusch & Ney (2009) and Albrecht & Hwa (2007a and 2007b); or focusing on the combination of a wide variety of metrics in a simple and straightforward way (Giménez 2008b, Specia and Giménez 2010). However, little research has been conducted on the effect of the use of linguistic features and how to combine them from a linguistic point of view. Therefore, our proposal is a linguistically-motivated metric which aims at using and combining varied linguistic knowledge at different levels in order to cover the key features that must be considered when dealing with MT evaluation from a linguistic point of view. Our hypothesis is that the use and combination of linguistic features at different levels will help us 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 design and on-going development of the VERTa metric. We detail the modules we are currently using and we report some preliminary experiments which will help us to discuss the use and combination of certain linguistic features in order to improve the metric performance. Moreover, for the sake of comparison, we also add a comparative study between VERTa and other well-known automatic metrics.

2. Methodology and metric design

When approaching the design and development of our linguistically-motivated metric, VERTa, we identified several linguistic issues which should be considered when comparing hypothesis and reference segments. These linguistic phenomena can be classified into lexical, phrase and clause level and they affected both syntax and semantics. Therefore, the linguistic knowledge that we intend to use is organised in different layers:

- **Lexical information:** At this level we want to highlight the importance of lexical semantics.

Lexical semantics becomes very important when using reference translations in order to evaluate MT output, because when using reference translations we cannot necessarily expect to find exactly the same word-forms in the hypothesis and the references. On the contrary, we must be able to establish lexical relations involving semantics such as synonymy, hyperonymy and hyponymy.

- **Morphological information:** Morphology is an important element, especially when dealing with languages with a rich inflectional morphology, such as Spanish, French and Catalan because it helps us to deal with linguistic features such as tense, aspect, mood, number, gender or case. Therefore, by means of morphological features such as tense, we can compare whether the tense used in the hypothesis and the reference translation is the same or varies. Moreover, inflectional morphology in combination with syntax (morphosyntax) also plays an important role in the sentence fluency. Such is the case of agreement in English, where verb forms in third person singular show agreement with the subject by means of the *–s* ending.
- **Syntactic information:** At this level a couple of issues are considered, the syntactic structure and the word order, both inside the phrase and inside the clause. In the syntactic structure we cover those changes that imply a change of grammatical category (i.e. passive-active alternation), and those that do not entail a change in the grammatical category of the units affected but account for the constituent word order. An example about the syntactic changes mentioned above is the following were the active-passive alternation is illustrated.

  Example 1:
  HYP: *...were assassinated by unknown men...*
  REF: *...unknown men assassinated...*

- **Sentence Semantics information:** This level is centred on sentence semantics and the causes which prevent a sentence from being partly or fully understood. Such is the case of the example below where the subject of the sentence realised by the proper noun *Merkel* is missing in the hypothesis sentences, and as a consequence we do not have information on the entity performing the action expressed by the verb.

  Example 2:
  HYP: *ø urged Tehran to...*
  REF: ***Merkel*** *called on Tehran...*

Therefore, our metric must account not only for a wide range of linguistic phenomena but also for different linguistic levels. In order to combine the above described linguistic features, we have decided to develop a metric which works at different stages, by means of organizing the linguistic information in several modules: lexical similarity metric, morphological similarity metric, dependency similarity and semantic similarity metric, respectively. Moreover, we have also added an n-gram similarity module so as to account for similarity between chunks. In addition, the organisation of linguistic features in different modules or levels allows us to evaluate both adequacy and fluency, thus trying to get closer to human evaluation. Given the stage of our work, currently we only focus on adequacy.

In this paper we describe the lexical, morphological, n-gram and dependency similarity metrics. The semantic similarity metric has not been explored so far, but we intend to do it in the near future. 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).

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. The metric is based on precision and recall and the traditional F-measure is applied in order to get the best final score for each pair of segments.

2.1 Lexical Similarity Module

Inspired by METEOR (Denkowsi and Lavie, 2011) the lexical similarity metric identifies matches between lexical items in the hypothesis segment and those in the reference segment taking into account several linguistic features. METEOR relies on the word-form, synonyms in

WordNet, stemming and paraphrasing. Our linguistic module uses some of these features (word-forms and synonyms) and adds others (hyperonyms, hyponyms, lemmas[1] and partial lemmas, i.e. lemmas that share the first 4 letters) in the established order (see Table 1). In addition, we also apply a system of weights (W) on the different matches which are manually established depending on their importance in terms of semantics.

W	Match	Examples	
		HYP	REF
1	Word-forms	east	east
1	Synonyms	believed	considered
.9	Direct-hyper.	Barrel	keg
.9	Direct-hypo.	Keg	barrel
.8	Lemma	is_BE	are_BE
.7	Partial-lemma	danger	dangerous

Table 1: Lexical matches and examples

2.2 Morphological Similarity Module

The morphological similarity metric combines lexical and morphological information in order to assess the fluency and well-formation of segments. This metric is based on the matches set in the lexical similarity metric, except for the partial-match, in combination with the Part of Speech (POS) tags from the annotated corpus[2]. By means of this combination, we can focus more on fluency and compensate the broader coverage that we have in the lexical module; therefore, preventing issues such as stating that *invited* and *invite* are positive matches regarding morphology. As a consequence, when assessing MT output in terms of fluency this metric will receive a higher weight, whereas when evaluating adequacy, the weight given to this module will be reduced. This module will be particularly useful when evaluating MT output of languages with a rich inflectional morphology, such as Romance languages.

Following the approach used in 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.

2.3 Dependency Similarity Module

The dependency similarity metric 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. By means of this module we are able to capture changes in word order and similarity between expressions which are comparable in their deep structure but different on their surface. This is illustrated in Example 3 where the adjunct of time *on Thursday*, although located differently in the hypothesis and reference segments, still depends on the verb *announced*, and therefore both segments show the same dependency analysis (Table 3).

Example 3
HYP: ... *Putin* ***on Thursday*** *announced that...*
REF: *Putin announced* ***on Thursday***..

HYPOTHESIS	REFERENCE
nsubj(Putin, announced)	nsubj(Putin, announced)
tmod(Thursday, announced)	tmod(Thursday, announced)

Table 3: Dependency analysis

Both hypothesis and reference strings are annotated with dependency relations by means of the Stanford parser (de Marneffe et al. 2006). The reason why this parser is used is because after conducting an evaluation (Comelles et al. 2010) where the performance of several dependency parsers was assessed (Stanford, DeSR, MALT, Minipar, RASP) this proved to be the best in terms of linguistic quality.

Similarly to the morphological module, the dependency similarity metric also relies first on the matches established at lexical level − word-form, synonymy, hyperonymy, 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, four different types of dependency matches have been designed as described below:

- **Complete** (CM): Type of match used when the triples are identical, this means that the label, the head and the modifier match.

  Label1(Head1,Mod1) = Label1(Head2,Mod2)
  Example 4:
  HYP: advmod(difficult, more)
  REF: advmod (difficult, more)

[1] Lemmas and lexical semantic relations are obtained by means of Wordnet 3.0.
[2] The corpus has been POS tagged using the Stanford Parser (de Marneffe et al. 2006).

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

Table 2: Morphological pairs of matches and examples

- **Partial** (PM): Three different types of partial matches are established:
    - Partial_no_mod (PM_no_mod): The label and the head match but the modifier does not match
        - Label1 = Label2
        - Head1 = Head2

      Example 5:
      HYP:**conj_and(difficult**, dangerous)
      REF: **conj_and(difficult**, serious)

    - Partial_no_head (PM_no_head): The label and the modifier match but the head does not match.
        - Label1 = Label2
        - Mod1 = Mod2

      Example 6:
      HYP: **prep_between**(mentioned, **Lebanon**)
      REF: **prep_between**(crisis, **Lebanon**)

    - Partial_no_label (PM_no_label): The head and the modifier match but the label does not match.
        - Head1 = Head2
        - Mod1 = Mod2

      Example 7:
      HYP: predet(**parties**, **all**)
      REF: det(**parties**,**all**)

Each type of match is given a weight which ranges from the highest to the lowest weight in the following order:

W	Dependency Match
1	Complete
.8	Partial_no_mod
.7	Partial_no_head
.7	Partial_no_label

Table 4: Weight for dependency matches

In addition, a couple of extra-rules have been added in order to capture the similarity between certain structures which are semantically equal but syntactically different. These structures are applied at phrase and sentence level. The former affects modifiers inside the noun phrase and the latter the passive-active voice relation. Regarding the structure inside the noun phrase, we cover the similarity between an adjective premodifiying a noun and an of-prepositional phrase postmodifying it, as exemplified below.

Example 8:
HYP: ...*between the* ***ministries of interior***...
REF: ...*between the two* ***interior ministries***...

The dependency triples obtained when analysing the NP in the hypothesis and reference strings, share the same head and modifier but they do not share the same label: for the NP in the hypothesis we get the triple *prep_of(ministries, interior)*, whereas for the NP in the reference we get *amod(ministries, interior)*. Although their labels differ, this couple of triples must be considered as an exact match due to their semantic similarity. Otherwise, we would penalise a couple of structures which are equal from a semantic point of view. At a clause level, structures used to express active-passive voice must be under consideration (see Example 9).

Example 9:
HYP: *After meeting* ***the Moroccan news agency published*** *a joint statement...*
REF: *A joint statement* ***published*** *(...)* ***by the Moroccan news agency****...*

In the hypothesis sentence, the dependency relation between the NP *Moroccan news agency* and the verb *published* is that of *nsubj(published, agency)*, whereas in the reference sentence the relation between the PP *by the Moroccan news agency* and the verb *published* is that of *agent(published, agency)*. Similar to the pair of dependencies dealing with modifiers, although the labels are different (*nsubj* and *agent*), both structures must be considered identical regarding their meaning and therefore, the previous couple of triples must be scored as an exact match.

2.4 N-gram Similarity Module

The n-gram similarity module is aimed at matching chunks[3] in the hypothesis and reference segments, taking as a starting point the matches obtained at lexical level. Chunks length goes from bigrams to sentence length. The

[3] By chunks we understand a group of words that go together, one next to the other.

n-gram similarity module uses the matches obtained at lexical level in order to align chunks. Thus, we do not only match n-grams relying on the word-form but also taking into account synonymy, hyponymy/hyperonymy and lemmas, as shown in example 10, where the chunks [*the situation in the area*] and [*the situation in the region*] match, although *area* and *region* do not share the same word-form but a relation of synonymy.

Example 10:
HYP: … the situation in the *area*…
REF: … the situation in the *region*…

2.5 Metrics Combination

As previously mentioned, VERTa aims at combining a varied range of linguistic features. Some approaches on combinining different metrics have been already explored, such as Leusch & Ney (2009), who combine several edit distance measures such as CDER and PER; Albrecht & Hwa (2007a and 2007b) who combine string-based and syntax based metrics by means of a regression-based learning approach; Giménez and Márquez (2008b), who combine a wide range of MT metrics by means of a non-parametric approach, they just choose the metric which performs best at each level and calculate the average score. However, very little research has been conducted to check the effect of each linguistic feature and how they should be combined depending on the type of evaluation, from a linguistic point of view. In this sense, VERTa combines different linguistic features by means of the combination of several modules which receive a specific weight depending on the type of evaluation: either adequacy or fluency. In the experiments reported below, adequacy has been evaluated and therefore, the lexical and dependency metrics receive higher weights than the morphology and n-gram similarity metrics. For the experiments reported in this paper weights have been set as follows:

- Lexical Module: 0.444
- Morphology Module: 0.111
- N-gram Module: 0.111
- Dependency Module: 0.333

3. Experiments

Some preliminary experiments have been conducted in order to obtain information on the suitability of the linguistic features used in our on-going metric. We wanted to focus our experiments on the use of hyperonyms and hyponyms as they affect several modules in our metric and have not been used by any other metric before. Moreover, we were also interested in comparing our linguistically-based metric with the standard metric BLEU[4] as well as with others which rely on linguistic features. The latter group comprises METEOR, and a set of measures developed by Giménez and Márquez (2007) based on shallow - SP-Op(*) - and dependency parsing – DP-Oc(*) and DP-Or(*). So as to perform these experiments we used part of the development data provided in the MetricsMaTr 2010 shared-task[5]. From the data provided by the organization we used 100 segments of the NIST Open-MT06 data, the MT output from 8 different MT systems (a total of 28,000 words approximately) and 4 reference translations. The human judgments used were based on adequacy (7-point scale, straight average). In order to calculate correlations at segment level Pearson correlation was applied between our metric and the adequacy judgments. All segments were taken into account regardless of the system providing them, in order to have a more precise correlation.
Our first experiment assesses the use of hyperonyms and hyponyms. We run our metric firstly with hyperonyms and hyponyms and secondly without them. Contrary to what we expected, the use of hyperonyms and hyponyms slightly weakened the metrics performance, as shown in Table 5 (where HYP comprises both hyperonyms and hyponyms).

	VERTa + HYP	**VERTa - No HYP**
Pearson Correlation	0.759	0.763

Table 5: Pearson correlation at segment level between VERTa and adequacy judgments

A close analysis of the data revealed that the use of hyperonyms and hyponyms had a positive effect when comparing some segments, such as those illustrated in Example 11, where thanks to this semantic relation our metric matches the words *press* and *papers*. However, we also found out that the use of these semantic relations introduced noise in our metric and this affected the way we applied the different matches set in each module. This is the case of Example 12 where the word *today* in the reference string shows a pair of matches: a) hyperonym-hyponym match between *day* & *today* and b) exact match between *today* & *today*.

Example 11:
HYP: … in protest against the publication of the prophet charges in European *papers* business…
REF: … in protest against the publication of caricatures of Prophet Muhamad in the European *press*…

Example 12:
HYP: … the situation in the area had not yet who is on its danger mark *day* **today**…
REF: …the situation in the region as having never been as deangerous as it is *today*…

Although the preferred match should be the exact match, our heuristics work from left to right and give priority to the hyperonym-hyponym match.

For the sake of comparison and so as to check that we were in the right direction, we compared VERTa to other metrics at segment level (see Table 6) by means of the earlier-mentioned adequacy judgments. We decided not to use only BLEU, one of the most used metrics, but also

[4] Despite the fact that BLEU does not necessarily measure the same kind of information.

[5] http://www.nist.gov/itl/iad/mig/metricsmatr10.cfm.

other metrics which are based on linguistic knowledge. We used METEOR-sy, which uses exact, synonym and stemming matching; METEOR-pa, which adds the paraphrase matching; SP-Op(*), based on lexical overlapping according to POS; DP-Oc(*) and DP-Or(*), based on dependency analysis.

Metric	Pearson Correlation
METEOR-pa	**0.766**
VERTa (no hyps)	**0.763**
METEOR-sy	0.744
BLEU	0.683
SP-Op(*)	0.622
DP-Oc(*)	0.408
DP-Or(*)	0.402

Table 6: Metric comparison at segment level

Interestingly, although being in its first stage, our metric outperforms n-gram-based metric BLEU, which already confirms that the use of rich linguistic knowledge is crucial in the evaluation process. Regarding the set of metrics based on shallow and dependency parsing, VERTa also gets a higher score. This is due to the fact that our metric does not only rely on shallow and dependency parsing, but we also use knowledge based on lexical semantics. Moreover, those metrics which only rely on shallow and dependency parsing are highly influenced by the performance of the parser used to obtain the linguistic analysis, which might not always be ideal. This may be a reason for their somehow contradictory results, where shallow-parsing based information seems to achieve better results than the more complex dependency-parsing based ones.

Finally, VERTa obtains higher scores than METEOR-sy, but the performance of our metric is slightly worse than that of METEOR-pa. This fact also seems to confirm two important criteria: a) the importance of using linguistic information in MT evaluation, showing that generally the more linguistic information, the higher the scores are, and b) the fact that the higher the number of translation references (as provided by METEOR-pa) the higher the probability is to find a translation match. These results encourage us to continue working on the use of linguistic features, and especially in refining the dependency module and the weights used in these first experiments.

4. Conclusion

In this paper we have described the first steps of a linguistically-motivated metric, VERTa, which aims at combining linguistic features at different levels. We have described the modules implemented so far and how linguistic features have been combined. We have also reported the results obtained in some experiments aimed at checking the use of certain linguistic features, i.e. hyperonyms and hyponyms. We have also compared VERTa to other well-known metrics such as BLEU and METEOR and other metrics based on shallow and dependency parsing. The results obtained in the correlation with human judgments show that the use of linguistic information is necessary in MT evaluation, and that the combination of linguistic features at different levels (lexical, morphological and syntactic) helps in getting results which correlate better with human judgments.

Although the results obtained in the experiments are just preliminary, they are extremely helpful to continue with our on-going research and help us to focus on those parts which need deep improvement and refinement. Moreover, the figures obtained by our primary metric implementation when compared to other well-known metrics show promising results for the combination and use of a wide variety of linguistic features.

In a near future, we plan to keep working on the development of the metric by exploring the use of other linguistic information (i.e. multi-words treatment and the use of semantic information at sentence level). In addition, we also expect to improve the metric performance by improving our dependency module (i.e. refining the type of dependency labels and matches to take into account) and continue working on the tuning of the weights used both inside the modules and in metrics combination. As regards the meta-evaluation of the metric, we will analyze the coverage of each level separately and we will evaluate our metric not only in terms of adequacy but also in terms of fluency, in order to establish which linguistic features play a more important role in each type of evaluation. Finally, we would also like to test the robustness of VERTa with other languages with richer inflectional morphology such as Spanish.

5. Acknowledgments

We are very grateful to LDC for kindly providing the development data used in the MetricsMaTr 2010 shared-task.

This work has been partially funded by the Spanish Government (projects KNOW2, TIN-2009-14715-C04-03, and Holopedia, TIN2010-21128-C02-02).

6. References

Albrecth, J. S., Hwa R. (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.

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

Callison-Burch, C., Osborne, M. and Koehn. P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the EACL 2006*, pages 249–256.

Comelles, E., Arranz, V. and Castellon, I. (2010). Constituency and Dependency Parsers Evaluation. SEPLN (ed.), *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Valencia:SEPLN, v. 45, p. 59-66. Valencia, Spain. ISSN: 1135-5948

Denkowski M. J., Lavie, A. (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)*, pages 85–91,

Edinburgh, Scotland, UK.

Giménez, J., Màrquez, L. (2007). Linguistic features for automatic evaluation of heterogeneous MT systems. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (ACL)*, pages 256-264, Prague, Czech Republic.

Giménez J., Màrquez, L. (2008). A smorgasbord of features for automatic MT evaluation in *Proceedings of the 3rd Workshop on Statistical Machine Translation (ACL)*, pages 195-198, Columbus, OH.

Gimenez, J., (2008). Empirical Machine Translation and its Evaluation. Doctoral Dissertation. UPC, Barcelona, Spain.

Giménez, J., Márquez, L. (2007). Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the ACL 2nd SMT Workshop*, Prague, Czech Republic pages 256–264.

He, Y., Du J., Way, A. and van Genabith, J. (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, pages 349-353.

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

Leusch, G., Ney, H. (2008). BLEUSP, INVWER, CDER: Three improved MT evaluation measures. In *NIST Metrics for Machine Translation 2008 Evaluation (MetricsMATR08)*, Waikiki, Honolulu, Hawaii, October 2008.

Liu, D., Hildea, D. (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, pages 25–32.

De Marneffe, M.C., MacCartney, B. and Manning, C.D. (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.

De Marneffe, M.C., Manning, C.D. (2008). The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, Manchester, UK.

Owczarzak, K., van Genabith, J. and Way, A.. (2007). Dependency-Based Automatic Evaluation for Machine Translation. In *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, Rochester, New York, pages 80– 87.

Owczarzak, K., van Genabith, J. and Way, A.. (2007). Labelled Dependencies in Machine Translation Evaluation in *Proceedings of the ACL Workshop on Statistical Machine Translation,* Prague, Czech Republic, pages 104– 111,.

Papineni, K., Roukos, S., Ward. T. and Zhu, W. (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, pages 311-318.

Specia, L., Giménez, J. (2010). Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. In the *9th Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.