VERTa: Facing the Multilingual Experience of a Linguistically-based MT Evaluation

E. Comelles§, J. Atserias‡, V. Arranz*, I. Castellón§, J. Sesé§
§University of Barcelona
Gran Via Corts Catalanes 585
08007 Barcelona, Spain
‡Yahoo Labs
Av. Diagonal, 177
08018 Barcelona, Spain
*ELDA/ELRA
9 rue des Cordelières
75013 Paris, France

E-mail: elicomelles@ub.edu, jordi@yahoo-inc.com, arranz@elda.org, icastellon@ub.edu, jordi.sese@gmail.com

Abstract

There are several MT metrics used to evaluate translation into Spanish, although most of them use partial or little linguistic information. In this paper we present the multilingual capability of VERTa, an automatic MT metric that combines linguistic information at lexical, morphological, syntactic and semantic level. In the experiments conducted we aim at identifying those linguistic features that prove the most effective to evaluate adequacy in Spanish segments. This linguistic information is tested both as independent modules (to observe what each type of feature provides) and in a combinatory fastion (where different kinds of information interact with each other). This allows us to extract the optimal combination. In addition we compare these linguistic features to those used in previous versions of VERTa aimed at evaluating adequacy for English segments. Finally, experiments show that VERTa can be easily adapted to other languages than English and that its collaborative approach correlates better with human judgements on adequacy than other well-known metrics.

Keywords: MT evaluation, automatic metric, linguistically-based metric, combination of metrics, Spanish

1. Introduction

In the last years several MT evaluation campaigns have been carried out (WMT'09, WMT'10, WMT'11, WMT'12 and WMT'13[1]) boosting the development of MT evaluation metrics not only for English but also for other languages (e.g. Spanish). Some of the metrics participating in these campaigns use lightweight linguistic information at a very specific level or no linguistic information at all (e.g., METEOR, Banerjee and Lavie, 2011; AMBER, Chen et al 2012; TerrorCat, Fishel et al. 2012; TESLA family of metrics, Dahlmeier et al. 2011; WMPF and MPF, Popovic 2011; ROSE, Song and Cohn 2011; ATEC, Wong and Kit 2010; BLEU, Papineni et al. 2001). In this paper we present the multilingual capability of VERTa, a metric which aims at using and combining a wide variety of linguistic features at lexical, morphological, syntactic and semantic level. We show that, although VERTa uses richer linguistic knowledge than previously mentioned metrics, it can be easily adapted to another language than English, such as Spanish, and that the results obtained outperform those of other well-known metrics.

2. The VERTa MT metric

VERTa is organised into different modules: *Lexical module*, covering information related to word-form, synonymy[2], hypernymy, hyponymy, lemma and partial lemma; *Morphological module*, combining lexical information and Part of Speech (PoS); *Dependency module*, using dependency analysis; *Ngram module*, combining ngrams with lexical semantics information; and *Semantic module*, using NEs, time expressions and sentiment analysis (see Comelles et al. 2012 for a more detailed information about VERTa's design). VERTa combines these linguistic features using different similarity metrics per each type of information. Each metric works first individually and uses a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triples, n-grams, etc). The final score is the Fmean of the weighted combination of the Precision and Recall of each metric. This way, the different modules can be weighted depending on their importance regarding the type of evaluation (fluency or adequacy) and language assessed.
When adapting VERTa to Spanish we took into account those linguistic characteristics that more sharply distinguish English from Spanish such as: the richer inflectional morphology that Spanish shows, the wider variety of spelling changes when creating new words belonging to the same family, as well as the more flexible

[1] http://www.statmt.org/

[2] Lemmas and lexical semantic relations are obtained by means of Wordnet.

word order in Spanish. Most of the metrics with best results for Spanish use information related to morphemes (such as ATEC and AMBER), combine this information with either PoS tagging (such as MPF and WMPF) or information regarding lemma and PoS (such as TerrorCat), and finally, others use information related to lemmas, PoS and synonyms, such as TESLA. However, none of those metrics use information regarding dependency analysis. We consider that this type of analysis is an appropriate method to assess those languages that show a flexible word order, as it allows for relating constituents in a sentence regardless of their position. In addition, this type of analysis captures similarity between expressions which are comparable in their deep structure but different on their surface, thus being especially useful when assessing adequacy.

The next sections describe the different adaptations done in order to deal with Spanish.

2.1 Lexical Similarity Module

The lexical similarity metric identifies matches between lexical items. Table 1 shows the features taken into account for Spanish, as well as the weights assigned to each match, in this case, the same weight. We disregarded partial lemma due to the Spanish wider variety of spelling changes in words belonging to the same family.

W	MATCH	EXAMPLES	
		HYP	REF
1	Word-form	plantas (*plants*)	plantas (*plants*)
1	Synonymy	prisión (*prison*)	cárcel (*jail*)
1	Direct Hypernynm	embarcación (*boat*)	barca (*rowboat*)
1	Direct Hyponym	barca (*rowboat*)	embarcación (*boat*)
1	Lemma	era_SER (*was_BE*, imperfect)	fue_SER (*was_BE*, preterite)

Table 1. Lexical matches and examples.

As regards synonyms, since no disambiguation is performed, all the possible synsets are taken into account in order to determine if a synonym relation is held between two words. Regarding hyperonym relations we use the most frequent sense of each one of the words.

2.2 Morphological Similarity Module

This metric is based on the matches set in the lexical similarity metric in combination with the PoS tags from the annotated corpus[3], as shown in Table 2. Each type of match was assigned the same weight.

W	MATCH	EXAMPLES	
		HYP	REF
1	Word-form, PoS	plantas, NCFP000	plantas NFCP000
1	Syn., PoS	prisión, NCFS000	cárcel, NCFS000
1	Hypernym, PoS	embarcación, NCFS000	barca, NCFS000
1	Hyponym, PoS	barca, NCFS000	embarcación, NCFS000
1	Lemma, PoS	era_(SER, VSII1S0)	era_(SER, VSII1S0)

Table 2. Morphological pairs of matches and examples.

In the Spanish version of VERTa, the Lemma-PoS match seems to be crucial as it avoids misleading matches such as verb forms *era (was - imperfect)* and *fue* (*was - preterite*) in the example below:

Example 1
SOURCE: *his success in the marathon **was** unexpected*
HYP: *su éxito en el marathon **era** inesperado*
REF: *su éxito en el maratón **fue** inesperado*

2.3 Dependency Similarity Module

This dependency similarity metric works at sentence level and follows the approach used in He et al. 2010 with some linguistic additions in order to adapt it to our metric combination. We used Freeling (Padró and Stanilovsky 2012) to obtain the dependency relations for Spanish. The dependency similarity metric relies first on the 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, four different types of dependency matches have been designed which were assigned the same weight, as shown in Table 3. In addition, dependency relations (i.e. nsubj, dobj, etc.) can also receive different weights depending on how informative they are.

W	TYPE OF MATCH	MATCH DESCRIPTION
1	Complete	Label1=Label2 Head1=Head2 Mod1=Mod2
1	Partial_no_label	Label1≠Label2 Head1=Head2 Mod1=Mod2
1	Partial_no_mod	Label1=Label2 Head1=Head2 Mod1≠Mod2
1	Partial_no_head	Label1=Label2 Head1≠Head2 Mod1=Mod2

Table 3. Dependency matches.

[3] The corpus has been PoS tagged using Freeling (Padró and Stanilovsky, 2012).

2.4 Ngram Similarity Module

The Ngram similarity module is aimed at matching chunks in the hypothesis and reference segments, taking as a starting point the matches obtained at lexical level, as shown in the example below. Chunks length may go from bigrams to sentence length.

Example 2

SOURCE: *the action of accomplishing something*

HYPOTHESIS: *De [la acción de lograr algo]*

REFERENCE: [*La acción de conseguir algo]*

In the example above, all words in the hypothesis segment can be matched with those in the reference segment, except for the preposition "de". Words "lograr" and "conseguir" can be matched thanks to the use of synonymy in the Lexical module.

3. Experiments and Results

Experiments conducted aimed at a) studying which linguistic features were the most appropriate to evaluate the adequacy of a segment in Spanish; b) exploring and finding the most effective combination of VERTa's modules to evaluate adequacy; c) comparing the linguistic information used to evaluate Spanish and English; and finally d) comparing VERTa to other well-known metrics. In order to perform these experiments part of a corpus developed in the KNOW-2[4] project was used. The data contains: 187 WordNet glosses that had been translated from English into Spanish by means of two different systems (Apertium [5] and Google Translator [6]), four reference translations and human judgements provided by two different judges. Experiments were performed at segment level and correlation with human judgements on adequacy was calculated by means of Pearson correlation. In the first three experiments each module was set up as explained in Section 2. Linguistic features used in the Lexical and Morphology modules were granted the same weight. Likewise, matches used in the Dependency module were given the same weight.

3.1 Influence of linguistic features

The aim of the first experiment was studying the influence of linguistic features used in each module. The first thing that must be noticed is that in the Lexical module the Partial-lemma match has not been used because as expected, the variability in Spanish spelling does not allow for a correct use of this match. This linguistic decision has been confirmed by the correlation obtained when this feature is included in the Lexical module, which slightly decreases its performance when both reference 1 and all 4 references are available (see Table 4). In addition, the use of hypernyms and hyponyms also seems to improve the performance of the Lexical module. However, this increase is just a tendency and more data would be needed in order to confirm the appropriateness of such feature.

As regards the Morphology module, the Lemma-PoS match slightly improves the correlation of this module. Even though in terms of correlation this is not a significant improvement either, it has a positive effect from a linguistic point of view as the use of this match prevents misleading matches such as that exemplified in section 2.2.

Regarding the Dependency module, all matches are used except for the no-head match, which does not correlate well with human judgements when reference 1 is used. This tendency is also confirmed when the 4 references are used: the omission of the no-head match has a strong positive impact in the correlation of this module. In addition, in the Dependency module, dependency relations are assigned a different weight, thus allowing us to distinguish between those relations which are considered more informative (i.e. subject-verb) and those less informative (i.e. determiner-noun). The most informative relations are assigned 1, whereas the least informative ones are assigned 0.5.

Module		Ref. 1	4 refs.
Lexical	Partial-lemma	0.49381	0.60663
	No Part.-lemma	**0.50198**	**0.63764**
	Hyper./Hypo.	**0.49381**	**0.63764**
	No Hyper./Hypo.	0.49138	0.63550
Morph.	Lemma-PoS	**0.47239**	**0.60070**
	No Lemma-PoS	0.47192	0.60049
Depend.	No-head match	0.43068	0.50617
	No No-head m.	**0.45889**	**0.62405**
	Dep. relations same weight	0.43068	0.59338
	Dep. relations different weight	**0.44099**	**0.62405**
Ngram	2gram-length	**0.39259**	**0.62856**
	Sentence-length	0.36977	0.53841

Table 4. Influence of linguistic features

Finally, the Ngram module shows a better performance when 2-gram length is used than when sentence-lenght grams are used. Longer ngrams are more appropriate to assess the grammaticality of a sentence, since the ommision of a word such as a determiner affects its grammaticality although it does not prevent the sentence from being understood, as shown in Example 3. In the hypothesis segment, the chunk "tiene ø servicio excelente" (*has excellent service*) is a disfluent chunk because the determiner "un" (*a*) is missing; however, the meaning of the sentence is not affected at all.

Example 3

SOURCE: *the performance of duties by a waiter or servant; "that restaurant has excellent service "*

HYP: *El rendimiento de deberes por un camarero o criado; "aquel restaurante tiene ø servicio excelente".*

[4] http://ixa.si.ehu.es/know2

[5] http://www.apertium.org/

[6] http://translate.google.com/

Shorter ngrams length seems, therefore, to be more appropriate when evaluating adequacy.

3.2 Combination of modules

Once the linguistic features for each module were analysed and set, our next step was to explore the combination of such features by combining VERTa's different modules. Table 5 shows the results of each module separately for experiments with one and four references, respectively. In both cases, the module that shows the best correlation is the Lexical module, thus confirming the undeniable fact that lexical semantics plays a key role when evaluating adequacy. The Dependency module also obtains similar correlations in both cases and occupies the third position in the ranking. This indicates that the Dependency module helps to compare two different syntactic structures which show the same meaning.

However, the Morphology and Ngrams module swap positions. The Morphology module had a significant influence when only one reference was available, since it obtained the second best performance. On the other hand, the Ngram module got a really low correlation. However, when 4 references are used, the second position is occupied by the Ngram module, whereas the Morphological module seems to be the least influential. It must be noticed, though that when 4 references are used, the performance of each module is closer in terms of correlation with human judgements than when just reference 1 is considered.

Module	Reference 1	4 references
Lexical M.	**0.50198**	**0.63764**
Morph. M.	0.47239	0.60070
Dependency M.	0.45888	0.62405
Ngram M.	0.39259	0.62856

Table 5. Correlations with human judgements per module, using ref. 1/ using 4 refs.

A thorough analysis of the data shows that the first reference used contains rather free translations, whereas the style of the other three references is closer to the hypothesis. An example of this different style illustrated in the example below, where references 2, 3 and 4 are closer to the hypothesis than reference 1.

Example 4
SOURCE: *the departure of a vessel from a port*
HYP: *La salida de un barco de un puerto.*
REF1: *Acción de zarpar una embarcación*
REF2: *La partida de un navío de un puerto*
REF3*: La partida de un barco desde un puerto*
REF4*: La partida de un barco del puerto*

Since VERTa uses similarity measures, it is clear that the preference when selecting a reference to compare the hypothesis with the 4 references available will be reference 2, 3 or 4 which are closer in style than reference 1. This also explains the increase in the performance of the Ngram module when 4 references are available. The Ngram module is based on the matches established by the Lexical module, thus, once lexical matches are set, the ngram similarity between the hypothesis and reference 2, 3 and 4 is closer than between the hypothesis and reference 1. In order to confirm this point, separate correlations were calculated for each reference. Table 6 shows that for each reference the Lexical module correlates better with human judgements than the rest of modules, highlighting again the importance of lexical semantics. The module that correlates worst with human judgements is the Morphology module, except for reference 1, where the Ngram module is the one that correlates the worst. As explained above, this is mainly due to the free translations in reference 1. In addition, the low correlation of the Morphology module in most of the references was expected, as this module seems more appropriate to deal with fluency issues. As regards the use of the Dependency module, it proves effective in most of the references.

Reference	Module	Correlation
1	Lexical Module	**0.50198**
	Morph. M.	0.47239
	Dependency M.	0.45888
	Ngram M.	0.39259
2	Lexical Module	**0.57360**
	Morph. M.	0.50208
	Dependency M.	0.56191
	Ngram M.	0.54847
3	Lexical Module	**0.52240**
	Morph. M.	0.47449
	Dependency M.	0.50874
	Ngram M.	0.50812
4	Lexical Module	**0.47799**
	Morph. M.	0.38931
	Dependency M.	0.44511
	Ngram M.	0.44700

Table 6. Pearson's correlation per module using each reference separately.

In order to make a final decision on the combination of VERTa's modules, all references were used. From a linguistic point of view and taking into account the type of evaluation and the characteristics of the language evaluated, those modules that seem to be the most appropriate were first the Lexical and Dependency module. The Lexical module accounts for semantics at word level because it uses synonymy and hypernymy/hyponymy relations. In addition, it must be noticed that dependency relations are an interface between syntax and semantics since they account for the internal relations in a sentence, moving away from its surface structure. Hence, the Dependency module looks like a good candidate to evaluate sentence semantics. As for the Ngram and Morphology modules, the Ngram module does not seem to play a key role when evaluating adequacy, although it is more important than the

Morphology module, since word order in a sentence has a stronger influence in meaning than inflectional morphology. Bearing all this in mind, module's weights were first assigned manually, following linguistic criteria. Later, in addition, in order to calculate an upper-bound for the weight tunning, all possible weight combinations were tuned automatically using a 0.01 step. The results obtained (see Table 7) confirmed our initial hypothesis that the highest weights should be assigned to the Lexical module and the Dependency module as they account for the meaning of the sentence, whereas the Ngram module and especially the Morphology module play a minor role when assessing adequacy in Spanish.

	MANUAL W.	AUTO. W.
Lexical Mod.	0.45	0.46
Morph. Mod.	0.05	0.03
Depend. Mod.	0.40	0.32
Ngram Mod.	0.10	0.19
CORREL.	**0.65963**	**0.66110**

Table 7. Correlations obtained when using manual and automatically tuned weights.

3.3 Spanish VERTa vs. English VERTa

In addition, we were also interested in comparing VERTa's performance when evaluating Spanish and its performance when evaluating English data. Results obtained for Spanish contrast with those obtained in Comelles et al. 2012, where manual tuning of VERTa for English showed that the weight assigned to the Morphology module had to be rather low. This was later confirmed by automatic tuning which concluded that only the Lexical, Dependency and Ngram modules should be taken into account to assess adequacy (see Table 8).

	MANUAL W.	AUTOM. W.
Lexical Mod.	0.44	0.51
Morph. Mod.	0.11	0
Depend. Mod.	0.33	0.45
Ngram Mod.	0.11	0.04
CORREL.	**0.763**	**0.780**

Table 8. VERTa's correlation for English data.

Although it is difficult to compare the data set used for Spanish and the one used for English, because their size and genres are very different, some preliminary conclusions can be drawn. First, the Lexical and Dependency modules are the most effective and appropriate ones to evaluate the adequacy of a segment both in English and Spanish. Second, the Ngram module should also be used but its influence in determining the adequacy of a segment is not crucial. Finally, automatically tuned weights confirmed that whereas in English the Morphology module does not prove effective to evaluate the adequacy, in Spanish it might be taken into account, although its role is less significant than the Lexical and Dependency module's. The reason why this module should be slighlty considered in Spanish but not in English is that Spanish shows a richer inflectional morphology than English, although its influence might be stronger if fluency was assessed.

3.4 Comparing VERTa with other MT metrics

Once experiments aimed at analysing the adequacy of linguistic features to evaluate adequacy in Spanish were conducted and discussed, the most natural step was to compare VERTa to other well-known metrics in order to evaluate the metric itself. Metrics used to compare VERTa were BLEU, METEOR-ex (only exact matching), METEOR-st (exact matching plus stemming) and METEOR-pa (exact matching, stemming and paraphrasing) and a set of linguistically-based metrics available in Asiya tool (Giménez and Márquez 2010; González et al. 2012). In this set of metrics, a couple of them use shallow parsing: SP-Op(*) calculates the average lexical overlap over PoS and SP-Oc(*) calculates the average lexical overlap over all chunk types. Others capture similarities between dependency trees in the hypothesis and reference segments and use the MALT v3.2 parser to analyse the segments. DPm-Ol(*) calculates overlapping between words hanging at all levels, DPm-Oc(*) calculates overlapping between grammatical categories, and finally, DPm-Or(*) calculates overlapping between grammatical relations. Finally, CP metrics compare similarities between constituent parse trees in the hypothesis and reference segments. The Charniak and Johnson (2005) Max-Ent reranking parser is used to obtain the constituent trees. CP-Op(*) calculates lexical overlap over PoS and CP-Oc(*) calculates lexical overlap according to the phrase constituent. Results obtained are shown in Table 9.

Metric	Pearson Correlation
VERTa	**0.66110**
METEOR-ex	0.60170
METEOR-st	0.61522
METEOR-pa	0.62127
BLEU	0.55514
SP-Op(*)	0.57700
SP-Oc(*)	0.56247
DPm-Ol(*)	0.42853
DPm-Oc(*)	0.56161
DPm-Or(*)	0.44837
CP-Op(*)	0.52463
CP-Oc(*)	0.56843

Table 9. Comparisson between VERTa and other well-known metrics.

Results obtained show that VERTa outperforms the rest of metrics, although the METEOR family also obtains good results, especially the version that uses paraphrasing. This indicates that when assessing adequacy the metric must be

flexible enough to account for lexical semantic relations and different ways to express the same meaning. Ngram-based metrics, such as BLEU, do not show a good correlation with human judgements, mainly because they are too rigid and account for word order, as a consequence, the omission of a single determiner is penalised. Linguistically-based metrics show a lower performance than VERTa, this is mainly due to the fact that they do not use any kind of information regarding lexical semantics, thus showing a lower flexibility than VERTa or METEOR. It is also noticeable the lower performance of the metric that uses information on dependency relations (DPm-Or(*)), which was expected to obtain a higher correlation with human judgements. Such a low performance might be due to the performance of the parser used for Spanish.

Correlations aside, data was also analysed in detail in order to compare VERTa's and METEOR-pa's performance. This analysis indicates that synonymy relations and the Dependency module play a key role when comparing both metrics and are the main reason why VERTa outperforms METEOR-pa, as illustrated by examples 5 and 6.

Example 5

SOURCE: *the performance of duties by a waiter or servant; "that restaurant has excellent service "*

HYP: *El rendimiento de deberes por un camarero o criado; "aquel restaurante tiene servicio excelente".*

REF: *Cumplimiento de la tarea de un camarero o un sirviente; "este restaurante tiene un servicio excelente"*

Despite not being a very natural sentence, the hypothesis segment conveys the meaning of the source segment. Synonymy helps in matching "deberes" and "tareas", as well as "criado" and "sirviente".

Example 6

SOURCE: *a failure to maintain a higher state*

HYP: *Un fracaso de mantener un estado más alto.*

REF: *Fracaso en el intento de mantener un estado superior*

The hypothesis segment communicates the meaning of the source segment, although it is slightly disfluent. In addition, the reference translation is rather free, since "en el intento de" has been added despite the fact that it does not appear in the source sentence. Fortunately, the Dependency module helps in maintaining the core meaning of the sentence and accounts for the relation of "fracaso" and "mantener" despite the addition of "en el intento".

4. Conclusions and future work

Experiments indicate that VERTa can be easily adapted to other languages than English, e.g. Spanish, and deal with different linguistic phenomena that are not present in English. In addition, despite the fact that the existence and quality of the different NLP analyzers for languages other than English could be an issue, this does not seem to be the case for Spanish, or at least, it does not seem to affect VERTa's performance.

Experiments have also shown that when evaluating adequacy for both Spanish and English, the Lexical and Dependency modules are the most effective ones, followed by the Ngram module. However, due to language particularities, namely Spanish richer inflectional morphology, the Morphology module should also be used when evaluating Spanish segments adequacy.

It has also been proved that VERTa gets better results than other well-known metrics, leading to the conclusion that a more collaborative approach that accounts for different aspects of language achieves a better correlation with human judgements, than those approaches that focus on rather partial aspects. Even when the reference translations are rather free, VERTa's results are better, mainly due to the help of the Dependency module and lexical semantics relation; in other words, thanks to the use of a more collaborative approach.

In the future we plan to use a larger corpus that will help us confirm the tendency in the use of linguistic features indicated by the experiments conducted in this paper. In addition, we would also like to focus and analyse the impact that the low grammatical quality of the analyzed text has on the performance of the automatic tools used.

5. Acknowledgements

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

6. References

Charniak, E., Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, pp. 173--180.

Chen, B., Kuhn, R. and Foster, G. (2012). Improving AMBER, an MT Evaluation Metric. *Proceedings of the 7th Workshop on Statistical Machine Translation*. Montreal, Canada, pp. 59--63.

Comelles, E., Atserias, J., Arranz, V. and Castellón. I. (2012). VERTa: Linguistic features in MT evaluation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey, pp. 3944—3950.

Dahlmeier, D., Liu, Ch. and Tou Ng, H. (2011). TESLA at WMT 2011: Translation Evaluation and Tunable Metric. In *Proceedings of the 6th Workshop on Machine Translation*. Edinburgh, Scotland, UK, pp. 78--84.

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*. Edinburgh, Scotland, UK, pp. 85--91.

Fishel, M., Sennrich, R., Popović, M. and Bojar. O.

(2012). TerrorCat: a Translation Error Categorization-based MT Quality Metric. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Montreal, Canada, pp. 64--70.

Giménez, J. (2008). *Empirical Machine Translation and its Evaluation*. Doctoral Dissertation. UPC.

Giménez, J., Márquez, Ll. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94.

González, M., Giménez, J. and Márquez, Ll. (2012). A Graphical Interface for MT Evaluation and Error Analysis. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju, Republic of Korea, pp. 139--144.

Hall, J., Nivre, J. (2008). A Dependency-Driven Parser for German Dependency and Constituency Representations. In *Proceedings of the ACL-08: HLT Workshop on Parsing German (PaGe08)*. Columbus, Ohio, USA, pp 47--54.

He, Y., Du, J., Way, A. and van Genabith, J. (2010). The DCU Dependency-based Metric in WMT-Metrics MATR 2010. In *Proceedings of teh 5th Workshop on Statistical Machine Translation*. Uppsala, Sweden, pp. 349--353.

Padró, Ll., Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey, pp. 2473—2479.

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

Popović., M. (2011). Morphemes and POS tags for n-gram based evaluation metrics. In *Proceedings of the 6th Workshop on Statistical Machine Translation*. Edinburgh, Scotland, pp. 104--107.

Song, X., Cohn, T. (2011). Regression and Ranking based Optimisation for Sentence Level Machine Translation Evaluation. In *Proceedings of the 6th Workshop on Statistical Machine Translation*. Edinburgh, Scotland, UK, pp. 123--129.

Wong, B. T-M. & Kit, Ch. (2010). ATEC automatic evaluation of machine translation via word choice and word order. *Machine Translation* 23(2): 141-151. Springer.