

The UPC Participation at the Metrics MATR Challenge 2008

Jesús Giménez and **Lluís Màrquez**
TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, E-08034, Barcelona
{jgimenez,lluism}@lsi.upc.edu

Abstract

This document describes the participation of the NLP Group from the Technical University of Catalonia (UPC) at the NIST 2008 “Metrics for Machine Translation” Challenge.

Our proposal is based on a rich set of metrics operating at different linguistic levels, based on different similarity assumptions. Syntactic and semantic metrics have proven very effective in the past, significantly outperforming standard metrics, which limit their scope to the lexical dimension. However, contrary to our expectations, meta-evaluation results over the development data provided by the Challenge organizers do not seem to fully corroborate these findings, at least with respect to metric performance at the segment level. Lexical metrics attain already high levels of correlation with human assessments both at the system and segment levels. We argue this result is not necessarily contradictory. A possible explanation could be found in a hypothetical lack of heterogeneity in the test bed, specially regarding system typology.

1 Introduction

In this paper, we advocate in favor of the use of heterogeneous linguistic information for the purpose of Machine Translation Evaluation. Our proposal is based on a rich set of individual metrics devoted to analyze complementary translation quality aspects at various linguistic levels (i.e., lexical, syntactic and semantic).

Scores conferred by individual metrics are then averaged into a single measure of quality. This simple combination scheme has been successfully tested over a number of evaluation tasks, including the Arabic-to-English and Chinese-to-English translation exercises at the NIST 2004 and 2005 Evaluation Campaigns (Giménez and Màrquez, 2008a),

and the translation of European Parliament Proceedings from several European languages into English (Giménez and Màrquez, 2008c).

The rest of the paper is organized as follows. Section 2 provides a brief description of the heterogeneous set of metrics utilized as well as the formal definition of the combination scheme. Then, Section 3 presents experimental results over the development set. Finally, in Section 4, we conclude by discussing the results and outlining future work.

All the metrics presented together with an implementation of the combination method have been made publicly available inside the ULC evaluation package, which may be freely downloaded¹.

2 Approach

As sketched in the introduction, we have compiled a representative set of metrics operating at different linguistic levels, whose scores are averaged into a single evaluation measure.

2.1 A Heterogeneous Metric Set

We have resorted to several existing metrics, and we have also developed new ones, all published in previous works (Giménez and Màrquez, 2007; Giménez and Màrquez, 2008b). Although based on different similarity assumptions, in all cases, translation quality is measured by comparing automatic translations against a set of human reference translations.

Below, we provide a brief description of the metrics employed grouped according to the linguistic level at which they operate:

- **Lexical Similarity**

We have considered several publicly available widely-used metrics either based on edit distance (e.g., TER)², lexical precision (e.g.,

¹<http://www.lsi.upc.edu/~nlp/IQMT>

²WER(Nießen et al., 2000) and PER(Tillmann et al., 1997) measures have not been included in the package nor in this study due to the lack of a public version.

BLEU, NIST), lexical recall (e.g., ROUGE), or on a balance between these latter two measures (e.g., GTM, METEOR). Below, we provide a short list.

- **BLEU** (Papineni et al., 2001). BLEU computes lexical precision among n -grams up to length 4. We use version ‘11b’ of the NIST MT evaluation kit³.
- **NIST** (Doddington, 2002). The NIST metric is an improved version of BLEU by the National Institute of Standards and Technology. The main difference with BLEU is in the way of averaging n -gram scores. While BLEU relies on a geometric mean, NIST performs an arithmetic mean. Also NIST takes into account n -grams up to length 5. In addition, NIST weights more heavily n -grams which occur less frequently, as an indicator of their higher informativeness. We use version ‘11b’ of the NIST MT evaluation kit for the computation of NIST scores.
- **GTM** (Melamed et al., 2003)⁴. GTM computes an F-measure balancing lexical precision and recall (Melamed et al., 2003). We use GTM version 1.4. Three variants, corresponding to different values of the e parameter controlling the reward for longer matchings ($e \in \{1, 2, 3\}$), are considered.
- **METEOR** (Banerjee and Lavie, 2005)⁵. METEOR computes an F-measure based on unigram alignment. METEOR also includes a fragmentation score which accounts for word ordering. Besides, it allows for considering stemming and synonymy lookup based on WordNet (Fellbaum, 1998). We use METEOR version 0.6. Four variants are considered:
 - * **METEOR_{exact}**, running ‘exact’ module.
 - * **METEOR_{stem}**, running ‘exact’ and ‘porter_stem’ modules, in that order. This variant considers morphological variations through stemming.
 - * **METEOR_{wnstem}**, running ‘exact’, ‘porter_stem’ and ‘wn_stem’ modules, in that order. This variant includes morphological variations obtained using WordNet.

- * **METEOR_{wnsyn}**, running ‘exact’, ‘porter_stem’, ‘wn_stem’ and ‘wn_synonymy’ modules, in that order. This variant performs a lookup for synonyms in WordNet.
- **ROUGE** (Lin and Och, 2004)⁶. ROUGE computes lexical recall among n -grams up to length 4. It also allows for considering stemming and discontinuous matchings (skip bigrams). We use ROUGE version 1.5.5. We consider morphological variations through stemming⁷. Four variants have been selected:
 - * **ROUGE_L**, longest common subsequence (LCS).
 - * **ROUGE_{S*}**, skip bigrams with no max-gap-length.
 - * **ROUGE_{SU*}**, skip bigrams with no max-gap-length, including unigrams.
 - * **ROUGE_w**, weighted longest common subsequence (WLCS) with weighting factor $w = 1.2$.
- **TER** (Snover et al., 2006)⁸. TER measures the amount of post-editing that a human would have to perform to change a system output so it exactly matches a reference translation. Possible edits include insertions, deletions, and substitutions of single words as well as shifts of word sequences. All edits have equal cost. We use $1 - \text{TER}$.
- **O_l** (Giménez and Màrquez, 2007). O_l is a short name for lexical overlapping. Automatic and reference translations are considered as unordered sets of lexical items. O_l is computed as the cardinality of the intersection of the two sets divided into the cardinality of their union.

• Shallow Syntactic Similarity (SP)

Metrics based on shallow parsing (*SP*) analyze similarities at the level of parts of speech (PoS), word lemmas, and base phrase chunks. Sentences are automatically annotated using the SVMTool (Giménez and Màrquez, 2004)⁹, Freeling (Carreras et al., 2004)¹⁰ and BIOS (Surdeanu et al., 2005)¹¹ linguistic processors. We

³<http://www.nist.gov/speech/tools/>

⁴<http://nlp.cs.nyu.edu/GTM/>

⁵<http://www.cs.cmu.edu/~alavie/METEOR/>

⁶<http://berouge.com>

⁷ROUGE options are ‘-z SPL -2 -1 -U -m -r 1000 -n 4 -w 1.2 -c 95 -d’.

⁸<http://www.cs.umd.edu/~snover/tercom/>

⁹<http://www.lsi.upc.edu/~nlp/SVMTool/>

¹⁰<http://garraf.epsevg.upc.es/freeling/>

¹¹<http://www.surdeanu.name/mihai/bios/>

instantiate lexical overlapping (O_l) over parts of speech and chunk types. The goal is to capture the proportion of lexical items correctly translated, according to their shallow syntactic realization. Two metrics have been defined:

- **SP- O_p -*** Average lexical overlapping over parts-of-speech.
- **SP- O_c -*** Average lexical overlapping over base phrase chunk types.

At a more abstract level, we use the NIST metric to compute accumulated/individual scores over sequences of:

- **SP-NIST $_l$** Lemmas.
- **SP-NIST $_p$** Parts-of-speech.
- **SP-NIST $_c$** Base phrase chunks.
- **SP-NIST $_{iob}$** Chunk IOB labels¹².

• Syntactic Similarity

We have grouped syntactic metrics in two families, according to the type of parsing performed, constituency parsing (CP), and dependency parsing (DP).

– On Constituency Parsing (CP)

CP metrics analyze similarities between constituency parse trees associated to automatic and reference translations. Constituency trees are obtained using the Charniak-Johnson’s Max-Ent reranking parser (Charniak and Johnson, 2005). Three types of metrics are defined:

- * **CP-STM** This metric corresponds to a variant of the syntactic tree matching (STM) metric presented by Liu and Gildea (2005), which considers subtrees up to length 4.
- * **CP- O_p -*** Similarly to ‘SP- O_p -*’, this metric computes average lexical overlapping over parts-of-speech, which are now consistent with the full parsing.
- * **CP- O_c -*** Average lexical overlapping over phrase constituents. The difference between these metric and ‘SP- O_c -*’ variant is in the phrase scope. In contrast to base phrase chunks, constituents allow for phrase embedding and overlapping.

– On Dependency Parsing (DP)

DP metrics capture similarities between dependency trees associated to automatic and

reference translations. Dependency trees are obtained using the MINIPAR parser (Lin, 1998)¹³. We use two types of metrics:

- * **DP-HWC** These metrics correspond to variants of the head-word chain matching (HWCM) metric presented by Liu and Gildea (2005) slightly modified so as to consider different head-word chain types:

- **DP-HWC $_w$** words.
- **DP-HWC $_c$** grammatical categories.
- **DP-HWC $_r$** grammatical relations.

In all cases only chains up to length 4 are considered.

- * **DP- $O_l|O_c|O_r$** These metrics correspond exactly to the LEVEL, GRAM and TREE metrics introduced by Amigó et al. (2006):

- **DP- O_l -*** Average overlapping between words according to the level of the dependency tree they hang at.
- **DP- O_c -*** Average overlapping between words *directly hanging* from terminal nodes (i.e. grammatical categories) of the same type.
- **DP- O_r -*** Average overlapping between words ruled by non-terminal nodes (i.e. grammatical relationships) of the same type.

• Shallow-Semantic Similarity

We have designed two families of metrics, *NE* and *SR*, which are intended to capture similarities over Named Entities (NEs) and Semantic Roles (SRs), respectively.

– On Named Entities (NE)

NE metrics analyze similarities between automatic and reference translations by comparing the NEs which occur in them. Sentences are automatically annotated using the BIOS package. We have defined two types of metrics:

- * **NE- O_e -*** Average lexical overlapping between NEs of the same type. This metric focus only on actual NEs. We use also another variant, ‘**NE- O_e -****’, which includes overlapping among items of type ‘O’ (i.e., Not-a-NE).
- * **NE- M_e -*** Average lexical matching between NEs of the same type.

¹²IOB labels are used to denote the position (Inside, Outside, or Beginning of a chunk) and, if applicable, the type of chunk.

¹³<http://www.cs.ualberta.ca/~lindek/minipar.htm>

– **On Semantic Roles (SR)**

SR metrics analyze similarities between automatic and reference translations by comparing the SRs (i.e., arguments and adjuncts) which occur in the predicates. Sentences are automatically annotated using the SwiRL package (Surdeanu and Turmo, 2005)¹⁴. This package requires at the input shallow parsed text enriched with NEs, which is obtained using BIOS.

- * **SR- O_r -★** Average lexical overlapping between SRs of the same type.
- * **SR- M_r -★** Average lexical matching between SRs of the same type.
- * **SR- O_r** This metric reflects ‘role overlapping’, i.e.. overlapping between semantic roles independently from their lexical realization.

We also consider a more restrictive version of these metrics (‘SR- M_{rv} -★’, ‘SR- O_{rv} -★’, and ‘SR- O_{rv} ’), which require SRs to be associated to the same verb.

• **Semantic Similarity**

We have entered the semantic level following the Discourse Representation Theory by Kamp (1981).

– **On Discourse Representations (DR)**

DR metrics analyze similarities between automatic and reference translations by comparing their respective discourse representation structures (DRSs), as provided by the the C&C Tools (Clark and Curran, 2004)¹⁵. DRSs are essentially a variation of first-order predicate calculus which can be seen as semantic trees. We have defined three different kinds of metrics:

- * **DR-STM** Average syntactic tree matching considering semantic subtrees up to length 4.
- * **DR- O_r -★** Average lexical overlapping between DRSs of the same type.
- * **DR- O_{rp} -★** Average morphosyntactic overlapping between DRSs of the same type.

A deeply detailed description of these metrics may be found in (Giménez, 2008).

¹⁴<http://www.surdeanu.name/mihai/swirl/>

¹⁵<http://svn.ask.it.usyd.edu.au/trac/candc/>

2.2 Improved Sentence-Level Behavior of Semantic Features

Metrics based on deep linguistic analysis rely on automatic processors, trained on out-domain data, which may be, thus, prone to error. In order to improve the recall of these metrics, we back off to lexical overlapping, O_i , but only in those cases when the linguistic processor is not able to produce any linguistic analysis (Giménez and Màrquez, 2008b). We have applied this technique only to SR and DR metric variants.

2.3 Uniform Linear Combinations

Integrating the scores conferred by different metrics into a single measure seems the most natural and direct way to improve over the individual quality of current metrics. A number of approaches have been already suggested (Corston-Oliver et al., 2001; Kulesza and Shieber, 2004; Quirk, 2004; Gamon et al., 2005; Amigó et al., 2005; Liu and Gildea, 2007; Albrecht and Hwa, 2007a; Albrecht and Hwa, 2007b; Paul et al., 2007).

In this work, we have followed a uniformly-averaged linear combination scheme (ULC), i.e., arithmetic mean of metric scores (Giménez and Màrquez, 2008a). This approach is similar to that of Liu and Gildea (2007) except that in our case the contribution of each metric to the overall score is not adjusted. Formally:

$$ULC_X(a, R) = \frac{1}{|X|} \sum_{x \in X} x(a, R)$$

where X is the metric set, and $x(a, R)$ is the similarity between the automatic translation a and the set of references R , for the given test case, according to the metric x . Let us note that all metrics used in this work are in a $(0, 1)$ range, with the only exception of NIST scores.

Optimal metric sets are determined by maximizing the correlation with human assessments at the segment level.

3 Experimental Work

In this section, we present meta-evaluation results over the development data provided by the Challenge organizers. We have considered several metric representatives from each linguistic level. We also report results on the combination of individual metric scores on the basis of the ULC scheme.

3.1 Settings

We have limited to the ‘mt06’ part of the ‘dev-set’, which corresponds to the NIST 2006 Open MT Eval-

	mt06
#references	4
#systems	8
#segments	249
Adequacy _{avg}	5.38/7

Table 1: Test bed description

uation Campaign. This set consist of a selection of 25 documents, totalling 249 segments. For the purpose of automatic evaluation, 4 human reference translations and automatic outputs by 8 different MT systems are available. In addition, we count on the results of a process of manual evaluation. All translation outputs have been evaluated in terms of adequacy on a 1-7 scale by a single human judge. A brief numerical description of this test bed is available in Table 1. Average adequacy is provided as an indicator of the overall translation quality exhibited by automatic systems.

In our experiments, metrics are evaluated in terms of human acceptability, as measured on the basis of correlation with human assessments both at the segment and system levels. Specifically, we compute Pearson correlation coefficients between metric scores and adequacy assessments.

3.2 Results

Table 2 presents meta-evaluation results. Metrics are grouped according to the linguistic level at which they operate (i.e., lexical, shallow syntactic, syntactic, shallow-semantic and semantic). For the sake of readability, we have selected a small set of representatives from each level. Metric quality is evaluated in terms of correlation with human assessments, both at the system level (R_{sys} , column 1) and segment level (R_{seg} , column 2).

3.2.1 Individual Behavior

At the system level, all metrics attain high correlation coefficients. The fact that lexical metrics exhibit such a fine performance is a clear indicator of the low heterogeneity of the development test bed, as to system typology. In other words, all systems seem to produce translations which share the sub-language (lexical choice and word order) represented by the set of human reference translations, possibly because they all belong to the statistical paradigm. Metrics operating at deeper linguistic levels, exhibit a fine performance as well.

At the segment level, however, all metrics suffer a significant quality decrease. Top scoring metrics are ‘NIST’ and ‘ROUGE_L’, with a correlation coefficient around 0.7. The drop in performance is more acute in

Level	Metric	R_{sys}	R_{seg}
Lexical	1-TER	0.95	0.58
	BLEU	0.94	0.59
	NIST	0.95	0.70
	GTM($e = 1$)	0.88	0.51
	GTM($e = 2$)	0.93	0.54
	GTM($e = 3$)	0.93	0.49
	O_l	0.96	0.66
	ROUGE _L	0.98	0.71
	ROUGE _{S*}	0.96	0.65
	ROUGE _{SU**}	0.96	0.67
	ROUGE _W	0.98	0.60
	METEOR _{exact}	0.97	0.62
	METEOR _{stem}	0.97	0.63
	METEOR _{wnstm}	0.97	0.63
METEOR _{wnsyn}	0.98	0.65	
Shallow Syntactic	SP- O_p -*	0.93	0.58
	SP- O_c -*	0.95	0.64
	SP-NIST _l	0.95	0.69
	SP-NIST _p	0.97	0.52
	SP-NIST _{iob}	0.97	0.40
	SP-NIST _c	0.97	0.29
Syntactic	CP- O_p -*	0.96	0.64
	CP- O_c -*	0.96	0.65
	CP-STM	0.97	0.59
	DP- O_l -*	0.96	0.56
	DP- O_c -*	0.97	0.60
	DP- O_r -*	0.99	0.62
	DP-HWC _w	0.97	0.37
	DP-HWC _c	0.98	0.39
DP-HWC _r	0.98	0.40	
Shallow Semantic	NE- M_e -*	0.93	0.44
	NE- O_e -*	0.93	0.63
	NE- O_e -**	0.95	0.64
	SR- M_r -*	0.96	0.37
	SR- O_r -*	0.99	0.49
	SR- O_r	0.98	0.26
	SR- M_{rv} -*	0.95	0.34
	SR- O_{rv} -*	0.99	0.41
SR- O_{rv}	0.99	0.33	
Semantic	DR- O_r -*	0.93	0.57
	DR- O_{rp} -*	0.93	0.48
	DR-STM	0.92	0.42
ULC _{opt}		0.98	0.74
ULC _h		0.98	0.65

Table 2: Meta-evaluation results based on human acceptability for the MATR 2008 Challenge development set

the case of linguistic metrics. This could be possibly due to a lack of robustness, and/or to the fact that these are partial measures of quality.

3.2.2 Collective Behavior

We have computed ULC over two different sets of metrics:

ULC_{opt} Based on the set of metrics of optimal¹⁶ correlation with adequacy at the segment level, over the development set.

$$M_{opt} = \{ \text{ROUGE}_L, \text{ROUGE}_W, \text{METEOR}_{w\text{nsyn}}, O_l, \text{DP-}O_{r-\star}, \text{DR-}O_{rp-\star} \}.$$

This combined set attains an improved segment-level evaluation quality (from 0.71 to 0.74). At, the system-level, its performance is comparable to the top scoring individual metrics, already close to 1.

ULC_h Employing a heuristically predefined set of metrics, including a couple of representatives from each linguistic level, based on our experience over different test beds (Giménez and Màrquez, 2007; Giménez and Màrquez, 2008a; Giménez and Màrquez, 2008b; Giménez and Màrquez, 2008c).

$$M_h = \{ \text{ROUGE}_W, \text{METEOR}_{w\text{nsyn}}, \text{CP-STM}, \text{DP-HWC}_c, \text{DP-HWC}_r, \text{DP-}O_{r-\star}, \text{SR-}O_{r-\star}, \text{SR-M}_{r-\star}, \text{SR-}O_r, \text{DR-}O_{r-\star}, \text{DR-}O_{rp-\star} \}.$$

Interestingly, this set, which does not include neither ‘NIST’ nor ‘ROUGE_L’, attains a reasonably high level of correlation, both at the segment level (0.65) and system level (0.98).

3.3 Our submission

Based on the previous results and also on our experience in the evaluation and meta-evaluation of a wide variety of evaluation scenarios, we decided to submit 6 different metrics. First, 2 combined metrics:

ULC_h The ULC combination over the heuristic set of metrics, M_h .

ULC_{opt} The ULC combination over the optimal set of metrics, M_{opt} .

Second, 4 individual metrics:

DP- $O_{r-\star}$ Average lexical overlapping over grammatical dependency relations.

SR- $O_{r-\star}$ Average lexical overlapping over semantic roles.

DR- $O_{r-\star}$ Average lexical overlapping over discourse representations.

DP- $O_{rp-\star}$ Average morphosyntactic overlapping over discourse representations.

The reason for sending this latter group of individual metrics is that we are interested in knowing their isolated performance over the, presumably heterogeneous, test set built by the Challenge organizers, and also in comparison to other metrics.

4 Conclusions and Future Work

We have presented a case study on the application of a heterogeneous set of metrics covering a wide variety of similarity aspects. Linguistic features allow system developers to separately analyze translation quality from complementary viewpoints, which, in its turn, allows for finer processes of error analysis (Giménez and Màrquez, 2008d). Although, linguistic features have been proven to lead to more reliable system rankings than metrics based on lexical matching alone (Giménez and Màrquez, 2007), this is not the case over the development data utilized in this work, possibly due to a low level of system heterogeneity.

The proposed combination scheme, although fairly simple, has been shown in the past to be an effective means of improving over the evaluation quality of individual metrics (Giménez and Màrquez, 2008a; Giménez and Màrquez, 2008c). Averaging scores produced by different metrics, ULC is indeed rewarding agreement between metrics. In other words, if different metrics, operating at different quality dimensions and based on different similarity assumptions, confer, in average, a high score to a given candidate translation, this translation is likely to be close to correctness, whereas if the average score is low, it is likely to be nonsense. Results over the development data provided by the Challenge organizers, however, do not reflect such clear improvements. We believe this behavior may be also attributable to a hypothetical lack of system heterogeneity in the test bed.

Based on this and other previous experiences over several test beds, we strongly believe that our metrics should be considered in future evaluation campaigns. For instance, the ULC_h metric could be applied. This variant has been shown to attain high levels of correlation with human assessments over several evaluation scenarios, outperforming in all cases most individual metrics, with the important advantage of

¹⁶Because exploring all possible metric combinations was not viable, we have used a simple algorithm which performs an approximate search. First, individual metrics are ranked according to their quality (according to R_{seg}). Then, following that order, metrics are added to the optimal set only if in doing so the global quality increases.

not requiring any adjustment of the heuristically defined metric set.

For future work, we are currently studying different alternative approaches to the combination of metric scores, such as the construction of human likeness classifiers, as suggested by Kulesza and Shieber (2004). We are also working on the design of metrics which truly operate at the document level, i.e., not by averaging sentence-level scores, but through the computation of similarities at the discourse level.

Acknowledgements

This research has been funded by the Spanish Ministry of Education and Science, project OpenMT (TIN2006-15307-C03-02). Our NLP group has been recognized as a Quality Research Group (2005 SGR-00130) by DURSI, the Research Department of the Catalan Government. We are grateful to the Metrics MATR Challenge organizers for providing such valuable test beds.

References

- Joshua Albrecht and Rebecca Hwa. 2007a. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 880–887.
- Joshua Albrecht and Rebecca Hwa. 2007b. Regression for Sentence-Level MT Evaluation with Pseudo References. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 296–303.
- Enrique Amigó, Julio Gonzalo, Anselmo Pe nas, and Felisa Verdejo. 2005. QARLA: a Framework for the Evaluation of Automatic Summarization. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. MT Evaluation: Human-Like vs. Human Acceptable. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.
- Xavier Carreras, Isaac Chao, Lluís Padró, and Muntxa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th LREC*, pages 239–242.
- Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 104–111.
- Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 140–147.
- George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology*, pages 138–145.
- C. Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.
- Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-Level MT evaluation without reference translations: beyond language modeling. In *Proceedings of EAMT*, pages 103–111.
- Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th LREC*, pages 43–46.
- Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 256–264.
- Jesús Giménez and Lluís Màrquez. 2008a. Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. In *Proceedings of IJCNLP*, pages 319–326.
- Jesús Giménez and Lluís Màrquez. 2008b. On the Robustness of Linguistic Features for Automatic MT Evaluation. To be published.
- Jesús Giménez and Lluís Màrquez. 2008c. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198.
- Jesús Giménez and Lluís Màrquez. 2008d. Towards Heterogeneous Automatic MT Error Analysis. In *Proceedings of the 6th LREC*.
- Jesús Giménez. 2008. *Empirical Machine Translation and its Evaluation*. Ph.D. thesis, Universitat Politècnica de Catalunya.
- Hans Kamp. 1981. A Theory of Truth and Semantic Representation. In J.A.G. Groenendijk, T.M.V. Janssen, , and M.B.J. Stokhof, editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematisch Centrum, address = Amsterdam.
- Alex Kulesza and Stuart M. Shieber. 2004. A Learning Approach to Improving Sentence-level MT Evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*.
- Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

- Dekang Lin. 1998. Dependency-based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*.
- Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.
- Ding Liu and Daniel Gildea. 2007. Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. In *Proceedings of the 2007 Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 41–48.
- I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.
- Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, RC22176. Technical report, IBM T.J. Watson Research Center.
- Michael Paul, Andrew Finch, and Eiichiro Sumita. 2007. Reducing Human Assessments of Machine Translation Quality to Binary Classifiers. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.
- Chris Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Metric. In *Proceedings of the 4th LREC*, pages 825–828.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231.
- Mihai Surdeanu and Jordi Turmo. 2005. Semantic Role Labeling Using Complete Syntactic Analysis. In *Proceedings of CoNLL Shared Task*.
- Mihai Surdeanu, Jordi Turmo, and Eli Comelles. 2005. Named Entity Recognition from Spontaneous Open-Domain Speech. In *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech)*.
- C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*.