IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation

Meritxell Gonzàlez, Alberto Barrón-Cedeño
TALP Research Center,
Technical University of Catalonia
{mgonzalez,albarron}@lsi.upc.edu

Lluís Màrquez
Qatar Computing Research Institute
Qatar Foundation
lmarquez@qf.org.qa

Abstract

This paper describes the UPC submissions to the *WMT14 Metrics Shared Task*: UPC-IPA and UPC-STOUT. These metrics use a collection of evaluation measures integrated in ASIYA, a toolkit for machine translation evaluation. In addition to some standard metrics, the two submissions take advantage of novel metrics that consider linguistic structures, lexical relationships, and semantics to compare both source and reference translation against the candidate translation. The new metrics are available for several target languages other than English. In the the official WMT14 evaluation, UPC-IPA and UPC-STOUT scored above the average in 7 out of 9 language pairs at the system level and 8 out of 9 at the segment level.

1 Introduction

Evaluating Machine Translation (MT) quality is a difficult task, in which even human experts may fail to achieve a high degree of agreement when assessing translations. Conducting manual evaluations is impractical during the development cycle of MT systems or for transation applications addressed to general users, such as online translation portals. Automatic evaluation measures bring valuable benefits in such situations. Compared to manual evaluation, automatic measures are cheap, more objective, and reusable across different test sets and domains.

Nonetheless, automatic metrics are far from perfection: when used in isolation, they tend to stress specific aspects of the translation quality and neglect others (particularly during tuning); they are often unable to capture little system improvements (enhancements in very specific aspects of the translation process); and they may make unfair comparisons when they are not able to reflect real differences among the quality of different MT systems (Giménez, 2008).

ASIYA, the core of our approach, is an open-source suite for automatic machine translation evaluation and output analysis.[1] It provides a rich set of heterogeneous metrics and tools to evaluate and analyse the quality of automatic translations. The ASIYA core toolkit was first released in 2009 (Giménez and Màrquez, 2010a) and has been continuously improved and extended since then (Gonzàlez et al., 2012; Gonzàlez et al., 2013).

In this paper we first describe the most recent enhancements to ASIYA: (*i*) linguistic-based metrics for French and German; (*ii*) an extended set of source-based metrics for English, Spanish, German, French, Russian, and Czech; and (*iii*) the integration of mechanisms to exploit the alignments between sources and translations. These enhancements are all available in ASIYA since version 3.0. We have used them to prepare the UPC submissions to the *WMT14 Metrics Task*: UPC-IPA and UPC-STOUT, which serve the purpose of testing their usefulness in a real comparative setting.

The rest of the paper is structured as follows. Section 2 describes the new reference-based metrics developed, including syntactic parsers for languages other than English. Section 3 gives the details of novel source-based metrics, developed for almost all the language pairs in this challenge. Section 4 explains our simple metrics combination strategy and analyses the results obtained with both approaches, UPC-IPA and UPC-STOUT, when applied to the WMT13 dataset. Finally, Section 5 summarises our main contributions.

2 Reference-based Metrics

We recently added a new set of metrics to ASIYA, which estimate the similarity between reference (*ref*) and candidate (*cand*) translations. The met-

[1] http://asiya.lsi.upc.edu

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 394–401,
Baltimore, Maryland USA, June 26–27, 2014.

rics rely either on structural linguistic information (Section 2.1), on a semantic mapping (Section 2.2), or on word n-grams (Section 2.3).

2.1 Parsing-based Metrics

Our initial set of parsing-based metrics is a follow-up of the proposal by Giménez and Màrquez (2010b): it leverages the structural information provided by linguistic processors to compute several similarity cues between two analyzed sentences. ASIYA includes plenty of metrics that capture syntactic and semantic aspects of a translation. New metrics based on linguistic structural information for French and German and upgraded versions of the parsers for English and Spanish are available since version 3.0.[2]

In the WMT14 evaluation, we opt for metrics based on shallow parsing (SP), constituency parsing (CP), and dependency parsing (DPm)[3]. Measures based on named entities (NE) and semantic roles (SR) were used to analyse translations into English as well. The nomenclature used below follows the same patterns as in the ASIYA's manual (Gonzàlez and Giménez, 2014). The manual describes every family of metrics in detail. Next, we briefly depict the concrete metrics involved in our submissions to the *WMT14 Shared Task*.

The set of SP metrics is available for English, German, French, Spanish and Catalan. They measure the lexical overlapping between parts-of-speech elements in the candidate and reference translations. For instance, SP-Op(VB) measures the proportion of correctly translated verbs; and the coarser SP-Op(*) averages the overlapping between the words for each part of speech. We also use NIST (Doddington, 2002) to compute accumulated scores over sequences of $n = 1..5$ parts of speech (SP-pNIST).

Similarly, CP metrics analyse similarities between constituent parse trees associated to candidate and reference translations. For instance, CP-STMi5 and CP-STM4 compute, respectively, the proportion of (individual) length-5 and accumulated up to length-4 matching sub-paths of the syntactic tree (Liu and Gildea, 2005). CP-Oc(*) computes the lexical overlap averaged over all the phrase constituents. Constituent trees are obtained using the parsers of Charniak and Johnson (2005), Bonsai v3.2 (Candito et al., 2010b), and Berkeley Parser (Petrov et al., 2006; Petrov and Klein, 2007) for English, French, and German, respectively.

Measures based on dependency parsing (DPm) — available for English and French thanks to the MALT parser (Nivre et al., 2007)— capture the similarities between dependency tree items (i.e., heads and modifiers). The pre-trained models for French were obtained from the French Treebank (Candito et al., 2010a) and used to train the Bonsai parser, which in turn uses the MALT parser. For instance, DPm-HWCM_w-3 retrieves average accumulated proportion of matching *word*-chains (Liu and Gildea, 2005) up to length 3; and DPm-HWCMi_c-3 computes the proportion of matching *category*-chains of length 3.

2.2 Explicit-Semantics Metric

Additionally, we borrowed a metric originally proposed in the field of Information Retrieval: explicit semantic analysis (ESA) (Gabrilovich and Markovitch, 2007). ESA is a similarity metric that relies on a large corpus of general knowledge to represent texts. Our knowledge corpora are composed of $\sim 100K$ Wikipedia articles from 2010 for the following target languages: English, French and German. In this case, *ref* and *cand* translations are both mapped onto the Wikipedia collection W. The similarities between each text and every article $a \in W$ are computed on the basis of the cosine measure in order to compose a similarities vector that represents the text. That is:

$$\vec{ref} = \{sim(ref, a)\ \forall a \in W\}\ , \quad (1)$$

$$\vec{cand} = \{sim(cand, a)\ \forall a \in W\}\ . \quad (2)$$

As the i-th elements in both $\vec{ref}$ and $\vec{cand}$ represent the similarity of *ref* and *cand* sentences to a common article, the similarity between *ref* and *cand* can be estimated by computing $sim(\vec{ref}, \vec{cand})$.

2.3 Language-Independent Resource-Free Metric

We consider a simple characterisation based on *word n-grams*. Texts are broken down into overlapping word sequences of length n, with 1-word shifting. The similarity between *cand* and *ref* is computed on the basis of the Jaccard coefficient (Jaccard, 1901). We used this metric for the pairs English–Russian and Russian-English, considering $n = 2$ (NGRAM-jacTok2ngram). For the

[2] Equivalent resources were previously available for English, Catalan, and Spanish.

[3] ASIYA includes two dependency parsers; the *m* identifies the metrics calculated using the MALT parser.

rest of the pairs we opt for the character-n-gram metrics described in Section 3.1, but they showed no positive results in the English–Russian pair during our tuning experiments.

3 Source-based Metrics

We enhance our evaluation module by including a set of new metrics that compare the source text against the translations. The metrics can be divided into two subsets: those that do not require any external resources (Section 3.1) and those that depend on a parallel corpus (Section 3.2).

3.1 Language-Independent Resource-Free Metrics

We opted for two characterisations that allow for the comparison of texts across languages without external resources nor language-related knowledge —as far as the languages use the same writing system.[4]

The first characterisation is *character n-grams*; proposed by McNamee and Mayfield (2004) for cross-language information retrieval between European languages. Texts are broken down into overlapping character sequences of length n, with 1-character shifting. We opt for case-folded bigrams (NGRAM-cosChar2ngrams), as they allowed for the best performance across all the pairs (except for *From/To* Russian pairs) during tuning.

The second characterisation (NGRAM-jacCognates) is based on the concept of *cognateness*; originally proposed for bitexts alignment (Simard et al., 1992). A word is a pseudo-cognate candidate if (*i*) it has only letters and $|w| \geq 4$, (*ii*) it contains at least one digit, or (*iii*) it is a single punctuation mark. *src* and *cand* sentences are then represented as word vectors, containing only those words fulfilling one of the previous conditions. In the case of (*i*) the word is cut down to its leading four characters only.

In both cases (*character n-grams* and *cognateness*) *cand* translations are compared against *src* sentences on the basis of the cosine similarity measure.

3.2 Parallel-Corpus Metrics

We consider two metrics that make use of parallel corpora: *length factor* and *alignment*.

[4]Previous research showed that transliteration is a good short-cut when dealing with different writing systems (Barrón-Cedeño et al., 2014).

Table 1: Length factor parameters as estimated on the WMT13 parallel corpora.

pair	μ	σ	**pair**	μ	σ
en–cs	0.972	0.245	*cs–en*	1.085	0.273
en–de	1.176	0.926	*de–en*	0.961	0.463
en–fr	1.158	0.411	*fr–en*	0.914	0.313
en–ru	1.157	0.678	*ru–en*	1.069	0.668

The length factor (LeM) is rooted in the fact that the length of a text and its translation tend to preserve a certain length correlation. For instance, translations from English into Spanish or French tend to be longer than their source. Similar measures were proposed during the statistical machine translation early days, both considering character- and word-level lengths (Gale and Church, 1993; Brown et al., 1991). Pouliquen et al. (2003) defines the length factor as:

$$\varrho(d') = e^{-0.5\left(\frac{\frac{|d'|}{|d_q|}-\mu}{\sigma}\right)^2}, \qquad (3)$$

where μ and σ represent the mean and standard deviation of the character lengths between translations of texts from L into L'. This is a stochastic normal distribution that results in higher values as the length of the target text approaches the expected value given the source. Table 1 includes the values for each language pair, as estimated on the WMT13 parallel corpora. Note that this metric was not applied to Hindi–English since this language pair was not present in the WMT13 challenge.

The last of our newly-added measures relies on the word alignments calculated between the sentence pairs *src–cand* and *src–ref*. We trained alignment models for each language pair using the Berkeley Aligner[5], and devised three variants of an ALGN metric, which compute: (*i*) the proportion of aligned words between *src* and *cand* (ALGNs); (*ii*) the proportion of aligned words between *cand* and *ref*, calculated as the combination of the alignments *src–cand* and *src–ref* (ALGNr); and (*iii*) the ratio of shared alignments between *src–cand* and *src–ref* (ALGNp).

4 Experimental Results

The tuning and selection of the different metrics to build UPC-IPA and UPC-STOUT was

[5]https://code.google.com/p/berkeleyaligner

conducted considering the *WMT13 Metrics Task* dataset (Macháček and Bojar, 2013), and the resources distributed for the *WMT13 Translation Task* (Bojar et al., 2013). Table 2 gives a complete list of these metrics grouped by families. First, we calculated the Pearson's correlation with the human judgements for all the metrics in the current version of the ASIYA repository, including standard MT evaluation metrics, such as METEOR (Denkowski and Lavie, 2011), GTM (Melamed et al., 2003), -TERp-A (Snover et al., 2009) (a variant of TER tuned towards adequacy), WER (Nießen et al., 2000) and PER (Tillmann et al., 1997). We selected the best performing metrics (i.e., those resulting in high Pearson coefficients) in each family across all the *From/To* English translation language pairs, including the newly developed measures —even if they performed poorly compared to others (see This is how the UPC-STOUT metrics sets for both *from* English and *To* English translation pairs were composed[6] (see Table 3).

Table 2: Metrics considered in the experiments separated by families according to the type of grammatical items they use.

1. -WER	17. DPm-HWCM_r-1
2. -PER	18. DPm-Or(*)
3. -TERp-A	19. SR-Or(*)
4. METEOR-ex	20. SR-Or
5. METEOR-pa	21. SR-Orv(*)
6. GTM-3	22. SR-Orv
7. SP-Op(*)	23. NE-Oe(*)
8. SP-pNIST	24. NE-Oe(**)
9. CP-STMi-5	25. ESA
10. CP-STMi-2	26. NGRAM-jacTok2ngrams
11. CP-STMi-3	27. NGRAM-jacCognates
12. CP-STM-4	28. NGRAM-cosChar2ngrams
13. CP-Oc(*)	29. LeM
14. DPm-HWCM_w-3	30. ALGNp
15. DPm-HWCM_c-3	31. ALGNs
16. DPm-HWCMi_c-3	32. ALGNr

Table 3: Metrics considered in each system.[7]

BAS: 1–6	SYN: 7–18
SEM: 19–25	SRC: 26–32
IPA: 1–9, 25–31	STOUT: 1–32

[6]Parser-based measures are not present in Czech nor Russian as target languages, ALGN is not available for French pairs, and ESA is not applied to Russian as target.

The metric sets included in UPC-IPA are light versions of the UPC-STOUT ones. They were composed following different criteria, depending on the translation direction. Parsing-based measures were already available in the previous version of ASIYA when translating into English — they are known to be robust across domains and are usually good indicators of translation quality (Giménez and Màrquez, 2007). So, in order to assess the gain achieved with these measures with respect the new ones, UPC-IPA neglects the measures based on structural information obtained from parsers. In contrast, this distinction was not suitable for the *From* English pairs since the number of resources and measures varies for each language. Hence, in this latter case, UPC-IPA used only the subset of measures from UPC-STOUT that required no or little resources.

In summary, when English is the *target* language, UPC-IPA uses the baseline evaluation metrics along with the length factor, alignments-based metrics, character n-grams, and ESA. In addition to the above metrics, UPC-STOUT uses the linguistic-based metrics over parsing trees, named entities, and semantic roles. When English is the *source* language, UPC-IPA relies on the basic collection of metrics and character n-grams only. UPC-STOUT includes the alignment-based metrics, length factor, ESA, and the syntactic parsers applied to both German and French.

In all cases (metric sets and language pairs), the translation quality score is computed as the uniformly-averaged linear combination (ULC) of all the individual metrics for each sentence in the testset. Its calculation implies the normalization of heterogeneous scores (some of them are negative or unbounded), into the range $[0, 1]$. As a consequence, the scores of UPC-IPA and UPC-STOUT constitute a natural way of ranking different translations, rather than an overall quality estimation measure. We opt for this linear combination for simplicity. The discussion below suggests that a more sophisticated method for weight tuning (e.g., relying on machine learning methods) would be required for each language pair, domain and/or task since different metric families perform notably different for each subtask.

We complete our experimentation by evaluating more configurations: BAS, a baseline

[7]These are the full sets of measures for each configuration. However, each specific subset for *From/To* English can vary slightly depending on the available resources.

Table 4: System-level Pearson correlation for automatic metrics over translations *From/To* English.

WMT13	*en–fr*	*en–de*	*en–es*	*en–cs*	*en–ru*	*fr–en*	*de–en*	*es–en*	*cs–en*	*ru–en*
UPC-IPA	93.079	85.147	88.702	85.259	70.345	96.755	94.660	95.065	94.316	72.083
UPC-STOUT	94.274	**90.193**	73.314	84.743	**70.544**	**96.916**	96.208	96.704	96.666	74.050
BAS	92.502	84.251	90.051	86.584	67.655	95.777	96.506	95.98	96.539	71.536
SYN	95.68	87.297	**96.965**	n/a	n/a	96.291	96.592	96.052	95.238	73.083
BAS+SYN	**94.584**	87.786	95.162	n/a	n/a	96.684	97.057	96.101	96.402	72.800
SEM	89.735	83.647	35.694	**95.067**	n/a	95.629	96.601	**98.021**	96.595	**76.158**
BAS+SEM	92.254	87.005	47.321	89.107	n/a	96.337	**97.534**	97.568	**97.371**	74.804
SRC	14.465	-16.796	-22.466	-49.981	39.527	13.405	-51.371	71.64	-73.254	68.766
BAS+SRC	93.637	76.401	83.754	64.742	54.128	95.395	90.889	93.299	89.216	71.882
WMT13-Best	94.745	93.813	96.446	86.036	81.194	98.379	97.789	99.171	83.734	94.768
WMT13-Worst	78.787	-45.461	87.677	69.151	61.075	95.118	92.239	79.957	60.918	82.058

Table 5: Segment-level Kendall's τ correlation for automatic metrics over translations *From/To* English.

WMT13	*en–fr*	*en–de*	*en–es*	*en–cs*	*en–ru*	*fr–en*	*de–en*	*es–en*	*cs–en*	*ru–en*
UPC-IPA	18.625	14.901	17.057	7.805	15.132	22.832	25.769	26.907	21.207	19.904
UPC-STOUT	**19.488**	15.012	17.166	**8.545**	15.279	23.090	27.117	26.848	21.332	19.100
BAS	19.477	13.589	16.975	8.449	**15.599**	**24.060**	**28.259**	**28.381**	**23.346**	**20.983**
SYN	16.554	14.970	16.444	n/a	n/a	22.365	24.289	23.889	20.232	17.679
BAS+SYN	19.112	**16.016**	**18.122**	n/a	n/a	23.940	28.068	27.988	23.180	19.659
SEM	12.184	9.249	10.871	3.808	n/a	17.282	19.083	20.859	15.186	14.971
BAS+SEM	19.167	13.291	15.857	7.732	n/a	22.024	25.788	26.360	21.427	19.117
SRC	2.745	2.481	1.152	1.992	5.247	2.181	1.154	8.700	-4.023	16.267
BAS+SRC	18.32	13.017	15.698	7.666	13.619	22.292	24.948	26.780	17.603	20.707
WMT13-Best	21.897	19.459	20.699	11.283	18.899	26.836	29.565	24.271	21.665	25.584
WMT13-Worst	16.753	13.910	3.024	4.431	13.166	14.008	14.542	14.494	9.667	13.178

with standard and commonly used MT metrics; SYN, the reference-based syntactic metrics; SEM, the reference-based semantic metrics; SRC, the source-based metrics; and the combination of BAS with every other configuration: BAS+SYN, BAS+SEM, and BAS+SRC. Their purpose is to evaluate the contribution of the newly developed sets of metrics with respect to the baseline. The composition of the different configurations is summarised in Tables 2 and 3.

Evaluation results are shown in Tables 4 and 5. For each configuration and language pair, we show the correlation coefficients obtained at the *system-* and the *segment-level*, respectively. As customary with the WMT13 dataset, Pearson correlation was computed at the system-level, whereas Kendall's τ was used to estimate segment-level rank correlations. Additionally to the two submitted and seven extra configurations, we include the coefficients obtained with the *Best* and *Worst* systems reported in the official WMT13 evaluation for each language pair.

Although the results of our two submitted systems are not radically different to each other, UPC-STOUT consistently outperforms UPC-IPA. The currently available version of ASIYA, including the new metrics, allows for a performance close to the top-performing evaluation measures in last year's challenge, even with our naïve combination strategy.

It is worth noting that no configuration behaves the same way throughout the different languages. In some cases (e.g., with the SRC configuration), the bad performance can be explained by the weaknesses of the necessary resources when computing certain metrics. When analysed in detail, the cause can be ascribed to different metric families in each case. As a result, it is clear that specific configurations are necessary for evaluating different languages and domains. We plan to approach these issues as part of our future work.

When looking at the system-level figures, one can observe that the SEM set allows for a considerable improvement over the baseline system. The further inclusion of the SYN set —when available—, tends to increase the quality of the estimations, mainly when English is the source language. These properties impact on some of the UPC-STOUT configurations. In contrast, when looking at the segment-level scores, while

Table 6: System-level Pearson correlation results in the WMT14 Metrics shared task

	en–fr	*en–de*		*en–cs*	*en–ru*
UPC-IPA	93.7	13.0		96.8	92.2
UPC-STOUT	93.8	14.8		93.8	92.1
WMT14-Best	95.9	19.8		98.8	94.2
WMT14-Worst	88.8	1.1		93.8	90.3
	fr–en	*de–en*	*hi–en*	*cs–en*	*ru–en*
UPC-IPA	96.6	89.4	91.5	82.4	80.0
UPC-STOUT	96.8	91.4	89.8	94.7	82.5
WMT14-Best	98.1	94.2	97.6	99.3	86.1
WMT14-Worst	94.5	76.0	41.1	74.1	-41.7

Table 7: Segment-level Kendall's τ correlation results in the WMT14 Metrics shared task

	en–fr	*en–de*		*en–cs*	*en–ru*
UPC-IPA	26.3	21.7		29.7	42.6
UPC-STOUT	27.8	22.4		28.1	42.5
WMT14-Best	29.7	25.8		34.4	44.0
WMT14-Worst	25.4	18.5		28.1	38.1
	fr–en	*de–en*	*hi–en*	*cs–en*	*ru–en*
UPC-IPA	41.2	34.1	36.7	27.4	32.4
UPC-STOUT	40.3	34.5	35.1	27.5	32.4
WMT14-Best	43.3	38.1	43.8	32.8	36.4
WMT14-Worst	31.1	22.5	23.7	18.7	21.2

the SYN measures still tend to provide some gain over the baseline, the SEM ones do not. Finally, it merits some attention the good results achieved by the baseline for translations into English. We may remark here that our baseline included also the best performing state-of-the-art metrics, including all the variants of METEOR, that reported good results in the WMT13 challenge.

Tables 6 and 7 show the official results obtained by UPC-IPA and UPC-STOUT in WMT14.[8] The best and worst figures for each language pair are included for comparison —the worst performing submission at segment level is neglected as it seems to be a dummy (Macháček and Bojar, 2014 to appear). Both UPC-IPA and UPC-STOUT configurations resulted in different performances depending on the language pair. UPC-STOUT scored above the average for all the language pairs except for *en–cs* at both system and segment level, and *en–ru* at system level. Although the evaluation results are not directly comparable to the WMT13 ones, one can note that the results were notably better for pairs that involved Czech and Russian, and worse for those that involved French and German at system level. Analysing the impact of the evaluation methods and building comparable results in order to address a study on configurations for different languages is part of our future work.

5 Conclusions

This paper describes the UPC submission to the WMT14 metrics for automatic machine translation evaluation task. The core of our evaluation system is ASIYA, a toolkit for MT evaluation. Besides the formerly available metrics in ASIYA, we experimented with new metrics for machine translation evaluation, with especial focus on translation from English into other languages.

As previous work on English as target language has proven, syntactic and semantic analysis can contribute positively to the evaluation of automatic translations. For this reason, we integrated a set of new metrics for different languages, aimed at evaluating a translation from different perspectives. Among the novelties, (*i*) new shallow metrics, borrowed from Information Retrieval, were included to compare the candidate translation against both the reference translation (monolingual comparison) and the source sentence (cross-language comparison), including explicit semantic analysis and the lexical-based characterisations character n-grams and pseudo-cognates; (*ii*) new parsers for other languages than English were applied to compare automatic and reference translation at the syntactic level; (*iii*) an experimental metric based on alignments; and (*iv*) a metric based on the correlation of the translations' expected lengths was included as well. Our preliminary experiments showed that the combination of these and standard MT evaluation metrics allows for a performance close to the best in last year's competition for some language pairs.

The new set of metrics is already available in the current version of the toolkit ASIYA v3.0 (Gonzàlez and Giménez, 2014). Our current efforts are focused on the exploitation of more sophisticated methods to combine the contributions of each metric, and the extension of the list of supported languages.

Acknowledgements

This work was funded by the Spanish Ministry of Education and Science (TACARDI project, TIN2012-38523-C02-00).

[8] At the time of submitting this paper, the evaluation results for WMT14 Metrics Task were provisional.

References

Alberto Barrón-Cedeño, Monica Lestari Paramita, Paul Clough, and Paolo Rosso. 2014. A Comparison of Approaches for Measuring Cross-Lingual Similarity of Wikipedia Articles. *Advances in Information Retrieval. Proceedings of the 36th European Conference on IR Research*, LNCS (8416):424–429. Springer-Verlag.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning Sentences in Parallel Corpora. In Douglas E. Appelt, editor, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991)*, pages 169–176, Berkeley, CA, USA. Association for Computational Linguistics.

Marie Candito, Benot Crabb, and Pascal Denis. 2010a. Statistical French dependency parsing: treebank conversion and first results. In *The seventh international conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. 2010b. Benchmarking of Statistical Dependency Parsers for French. In *Proc. 23rd Intl. COLING Conference on Computational Linguistics: Poster Volume*, pages 108–116, Beijing, China.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-Fine N-best Parsing and MaxEnt Discriminative Reranking. In *Proc. 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 85–91, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

E. Gabrilovich and S. Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, San Francisco, CA, USA.

William A. Gale and Kenneth, W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19:75–102.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proc. of 2nd Workshop on statistical Machine Translation (WMT07), ACL'07, Prague, Czech Republic*.

Jesús Giménez and Lluís Màrquez. 2010a. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

Jesús Giménez and Lluís Màrquez. 2010b. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3–4):77–86.

Jesús Giménez. 2008. *Empirical Machine Translation and its Evaluation*. Ph.D. thesis, Universitat Politcnica de Catalunya, July.

Meritxell Gonzàlez and Jesús Giménez. 2014. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation, v3.0, February. http://asiya.lsi.upc.edu.

Meritxell Gonzàlez, Jesús Giménez, and Lluís Màrquez. 2012. A Graphical Interface for MT Evaluation and Error Analysis. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL). System Demonstration*, pages 139–144, Jeju, South Korea, July. Association for Computational Linguistics.

Meritxell Gonzàlez, Laura Mascarell, and Lluís Màrquez. 2013. tSearch: Flexible and Fast Search over Automatic translation for Improved Quality/Error Analysis. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics (ACL). System Demonstration*, pages 181–186, Sofia, Bulgaria, August.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

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*, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2014 (to appear). Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimare, US, June. Association for Computational Linguistics.

Paul McNamee and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2):73–97.

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)*, pages 61–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 (LREC)*.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. In *Proc. Human Language Technologies (HLT)*, pages 404–411. Association for Computational Linguistics, April.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. 2003. Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 401–408, Borovets, Bulgaria.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, pages 259–268, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*, pages 2667–2670.