IQ$_{MT}$: A Framework for Automatic Machine Translation Evaluation

Jesús Giménez* **and Enrique Amigó†**

*TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3. 08034, Barcelona
†Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Juan del Rosal, 16. 28040, Madrid
jgimenez@lsi.upc.edu, enrique@lsi.uned.es

Abstract

We present the IQ$_{MT}$ Framework for Machine Translation Evaluation Inside QARLA. IQ$_{MT}$ offers a common workbench in which evaluation metrics can be utilized and combined. It provides i) a measure to evaluate the quality of any set of similarity metrics (KING), ii) a measure to evaluate the quality of a translation using a set of similarity metrics (QUEEN), and iii) a measure to evaluate the reliability of a test set (JACK). The first release of the IQ$_{MT}$ package is freely available for public use. Current version includes a set of 26 metrics from 7 different well-known metric families, and allows the user to supply its own metrics. For future releases, we are working on the design of new metrics that are able to capture linguistic aspects of translation beyond lexical ones.

1. Introduction

In the last years, it has been repeatedly argued that current Machine Translation (MT) evaluation metrics do not capture well possible improvements attained by means of incorporating linguistic knowledge to MT systems (Och et al., 2003). One of the possible reasons is that most of the current metrics do not take into account any information at linguistic levels further than lexical. This is the case of metrics such as BLEU (Papineni et al., 2001), NIST (Doddington, 2002), WER (Nießen et al., 2000), PER (Leusch et al., 2003), GTM (Melamed et al., 2003), ROUGE (Lin and Och, 2004a), and METEOR (Banerjee and Lavie, 2005).

More sophisticated metrics are required. However, little work has been done in that direction. For instance, metrics such as ROUGE and METEOR may consider stemming. We may also find the WNM metric (Babych and Hartley, 2004), a variant of BLEU which weights n-grams according to their statistical salience estimated out from a monolingual corpus. Additionally, METEOR may perform a lookup for synonymy in WordNet (Fellbaum, 1998). But all these are still attempts at the lexical level. To our knowledge, the only attempt so far to exploit information at an upper level has been done by Liu and Gildea (2005) who introduced a series of syntax-based features based on syntactic tree matching.

Doubtless the design of a metric that is able to capture all the linguistic aspects that distinguish 'correct' translations from 'incorrect' ones is an ambitious and difficult goal. Instead of building such a sophisticated metric we suggest to follow a 'divide and conquer' strategy, and design a set of specialized metrics, devoted to the evaluation of partial aspects of MT quality. The new challenge is how to combine their outputs into a single measure.

In a recent work, Kulesza and Shieber (2004) tried to combine some aspects of different metrics by applying machine learning techniques to build classifiers that distinguished between human-generated ('good') and machine-generated ('bad') translations. They used features inspired in metrics like BLEU, NIST, WER and PER.

Our approach is based on QARLA (Amigó et al., 2005), a probabilistic framework originally designed for the evaluation of text summarization systems. QARLA automatically identifies the features that distinguish human translations from automatic ones. It permits metric combinations, without any a-priori weighting of their relative importance. Besides, no training or adjustment of parameters is required, there is no need for human assessments, and it does not depend on the scale properties of the metrics being evaluated. The methodology which is closest to QARLA is ORANGE (Lin and Och, 2004b). However, ORANGE does not permit metric combinations.

Our initial experiments applying QARLA to MT evaluation are discussed in (Giménez et al., 2005). As a result of our experience, in this work we present the IQ$_{MT}$[1] Framework for MT Evaluation, a public workbench in which similarity metrics may be robustly combined.

The rest of the paper is organized as follows. Section 2 describes the fundamentals of QARLA. The IQ$_{MT}$ architecture is deployed in Section 3. Section 4 presents a case of study on the evaluation of the Europarl Corpus Spanish-to-English translation task. Finally, ongoing work is outlined in Section 5.

2. Fundamentals

IQ$_{MT}$ is based on the QARLA Framework (Amigó et al., 2005). QARLA uses similarity to models (human references) as a building block. The main assumption is that all human references are equally optimal and, while they are likely to be different, the best similarity metric is the one

[1]The IQ$_{MT}$ Framework is publically available, released under the GNU Lesser General Public License (LGPL) of the Free Software Foundation. It may be freely downloaded at http://www.lsi.upc.edu/˜nlp/IQMT. Discussion on this software as well as information about oncoming updates takes place on the IQ$_{MT}$ google group, to which you can subscribe at http://groups-beta.google.com/group/IQMT.

that identifies and uses the features that are common to all human references, grouping and separating them from automatic translations.
Therefore, one of the main characterisitics of QARLA that differentiates it from other approaches, is that, besides considering the similarity of automatic translations to human references, QARLA additionally considers the distribution of similarities among human references.
The input for QARLA is a set of test cases A, a set of similarity metrics X, and sets of models R for each test case. With such a testbed, QARLA provides three measures:

- **KING**$_{A,R}(X)$, a measure to evaluate the descriptive power of a set of similarity metrics.
- **QUEEN**$_{X,R}(A)$**,** a measure to evaluate the quality of a translation using a set of similarity metrics.
- **JACK**(A, R, X)**,** a measure to evaluate the reliability of a test set.

QUEEN

QUEEN operates under the assumption that a good translation must be similar to all human references according to all metrics. QUEEN is defined as the probability, over $R \times R \times R$, that for every metric in X the automatic translation a is closer to a model than two other models to each other:

$$\text{QUEEN}_{X,R}(a) = Prob(\forall x \in X : x(a, r) \geq x(r', r''))$$

where a is the automatic translation being evaluated, $\langle r, r', r'' \rangle$ are three human references in R, and $x(a, r)$ stands for the similarity of r to a according to the similarity metric x. We can think of the QUEEN measure as using a set of tests (every similarity metric in X) to test the hypothesis that a given translation a is a model. Given $\langle a, r, r', r'' \rangle$, we test $x(a, r) \geq x(r', r'')$ for each metric x. a is accepted as a model only if it passes the test for every metric. Thus, QUEEN(a) is the probability of acceptance for a in the sample space $R \times R \times R$. This measure has some interesting properties:

(i) it is able to combine different similarity metrics into a single evaluation measure.

(ii) it is not affected by the scale properties of individual metrics, i.e. it does not require metric normalisation and it is not affected by metric weighting.

(iii) Peers (automatic translations) which are very far from the set of models (human references) all receive QUEEN=0. In other words, QUEEN does not distinguish between very poor translation strategies.

(iv) The value of QUEEN is maximised for peers that "merge" with the models under all metrics in X.

(v) The universal quantifier on the metric parameter x implies that adding redundant metrics does not bias the result of QUEEN.

However, the main drawback of QUEEN is that it requires the use of multiple references (at least three), when in most cases only a single reference translation is available.

KING

Based on QUEEN, QARLA provides a mechanism to determine the quality of a set of metrics, the KING measure:

$$\text{KING}_{A,R}(X) = Prob(\forall a \in A : \\ \text{QUEEN}_{X,R-\{r\}}(r) \geq \text{QUEEN}_{X,R-\{r\})}(a)$$

KING represents the probability that, for a given set of human references R, and a set of metrics X, the QUEEN quality of a human reference is greater than the QUEEN quality of *any* automatic translation in A. Therefore, KING measures the ability of a set of metrics to discern between automatic and human translations.

JACK

Again based on QUEEN, QARLA provides a mechanism to determine the reliability of the test set, the JACK measure:

$$\text{JACK}(A, R, X) = Prob(\exists a, a' \in A : \\ \text{QUEEN}_{X,R}(a) > 0 \wedge \text{QUEEN}_{X,R}(a') > 0 \\ \wedge \forall x \in X x(a, a') \leq x(a, r)$$

i.e. the probability over all human references r of finding a couple of automatic translations a, a' which are (i) close to all human references (QUEEN > 0) and (ii) closer to r than to each other, according to all metrics. JACK measures the heterogeneity of system outputs with respect to human references. A high JACK value means that most references are closely and heterogeneously surrounded by automatic translations. Thus, it ensures that R and A are not biased.

3. System Architecture

A schematic plot of the system architecture may be seen in Figure 1. IQMT consists of two main components, namely IQ*setup* and IQ*eval*. The IQ*setup* component is responsible for applying a set of similarity metrics to a set of automatic translations and a set of human references. The IQ*eval* component computes the KING, QUEEN, and JACK measures on top of the similarity scores generated by IQ*setup*.

3.1. IQ*setup*

IQ*setup* computes the similarities required for the estimation of the QUEEN measure. This component receives as input a configuration file specifying:

- set of human references (R)
- set of system outputs (A)
- set of metrics (X)

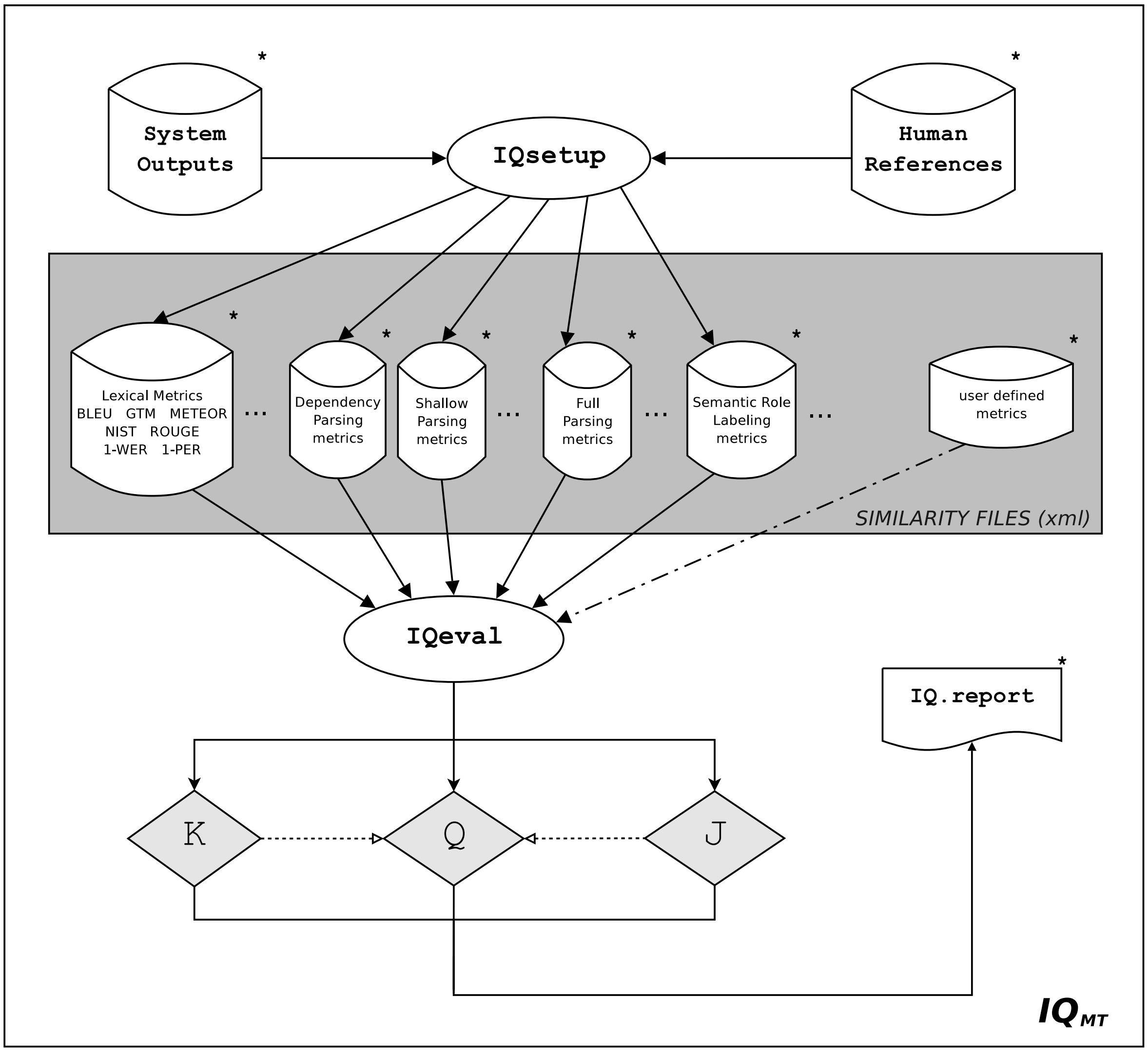

Figure 1: IQ$_{\text{MT}}$ system architecture.

Based on this information, IQ*setup* generates for each metric a collection of *'IQ XML'* similarity files:

- <metric>.<system-system>.xml
- <metric>.<system-reference+>.xml
- <metric>.<reference-reference+>.xml

These files are based on the *'IQ XML'* representation schema. See an example:

```
<IQ metric="BLEU-4" ref="R0"
          score="0.3945" target="S0">
  <S n="1">0.3033</S>
  <S n="2">0.5833</S>
  ...
  <S n="1007">0.6852</S>
  <S n="1008">0.8333</S>
</IQ>
```

The file above provides system and sentence level similarity scores obtained by comparing system 'S0' against reference 'R0' based on the 'BLEU-4' similarity metric.

The set of similarity metrics is a dynamic component in our framework. We have started by addapting existing MT evaluation metrics. These metrics are transformed into similarity metrics by considering just a single reference when computing its value. Current version integrates 26 variants from 7 families of metrics (BLEU, NIST, WER and PER, GTM, ROUGE, and METEOR)[2]. New metrics will be added in the future.
But the main advantage of the *'IQ XML'* representation schema is that it allows users to supply their own metrics in a transparent and unified manner. For every new metric, the user is responsible for generating an IQ XML similarity file for each pair <system-reference+>, <reference-reference+>, and <system-system>.

3.2. IQ*eval*

IQ*eval* allows us to calculate the KING, QUEEN, and JACK measures. Several options must be specified:

- the set of references
  (all references are used by default).

[2] A detailed list of the variants incorporated may be found in (Giménez et al., 2005)

- the set of system outputs to evaluate
  (all systems are evaluated by default).
- the set of metrics
  (all metrics are considered by default).

This information is specified according to a configuration file. The IQ*setup* component generates, as a by-pass product, a default configuration file for IQ*eval*. This configuration file contains a series of predefined sets. It must be edited in order to define new sets.
Other options are available:

- the subset of sentences per system to evaluate
  (all sentences are considered by default).
- level of granularity (sentence/system)
  (system level results are printed by default).

IQ*eval* also permits to obtain individual metric scores outside QARLA.

3.3. Finding an Optimal Metric Set

The optimal set is defined by the combination of metrics exhibiting the highest KING value. However, exploring all possible combinations is not viable[3]. IQ*eval* provides an implementation of a simple algorithm which performs an approximate search in order to find a suboptimal set of metrics:

1. Individual metrics are ranked by their KING value.
2. Following that order, metrics are individually added to the set of optimal metrics only if the global KING increases.

4. A case of study: Europarl

For a robust estimation of the KING, QUEEN, and JACK probabilities, the ideal scenario would consist of a large number of human references per sentence, and automatic outputs generated by heterogeneous MT systems. Unfortunately, this kind scenario is rarely found. Generally, few references are available (one in most cases), and MT systems are similar to each other. Thus, we have tested our system under a more realistic scenario. We utilize the data from the *'Openlab 2006'* Initiative[4] promoted by the TC-STAR[5] Consortium.

4.1. Experimental Setting

'Openlab 2006' data are entirely based on European Parliament Proceedings[6], covering April 1996 to May 2005. We focus on the Spanish-to-English translation task. The training set consists of 1,281,427 parallel sentences. For evaluation purposes we use the development set which consists of 1,008 sentences. Three human references per sentence are available. We intend to evaluate 4 systems:

- Word-based SMT system (WB).
- Systran Rule-based translation engine (SYSTRAN).
- Phrase-based SMT system (PB).
- Phrase-based SMT system (PB++)[7].

SMT systems are built as described in (Giménez and Màrquez, 2005). As to 'SYSTRAN', we used the freely available on-line version[8]. Let us note that evaluation is unfair to 'SYSTRAN' because SMT systems have been trained using in-domain data. However, we include 'SYSTRAN' for the sake of heterogeneity. We use all the 26 currently available metric variants.

4.2. Evaluating with Standard Metrics

First we analyze the individual behaviour of standard metrics. We use one representative from each family, the metric variant with highest KING value in the given test set.
See, in Table 1, the poor level of agreement between metrics. For instance, according to '1-PER' and '1-WER' the word-based SMT system (*'WB'*) is best. However, according to the rest of metrics the phrase-based systems (*'PB'* and *'PB++'*) are best, obtaining very similar scores. Also note that, contrary to our expectations, the 'SYSTRAN' system outperforms the word-based system according to five metrics and the two phrase-based systems according to two metrics. Therefore, the key question is "which metric should I trust?".

4.3. Evaluating with IQ_{MT}

Inside the IQMT Framework systems are evaluated according to their human-likeness. Thus, we must trust the metric (or set of metrics) with highest descriptive power (highest KING), i.e. the metric which best identifies the features that distinguish between human translations and automatic translations. We apply the algorithm described in Subsection 3.3. In the case of the *'Openlab 2006'* data, we can count only on three human references per sentence. In order to increase the number of samples for QUEEN estimation we can use reference similarities $x(r', r'')$ between manual translation pairs from other sentences, assuming that the distances between manual references are relatively stable across examples. The optimal set is:

{NIST-2, NIST-3, NIST-4, and 1-WER}

It attains a KING measure of 0.38, which means that in 38% of the cases this metric set is able to identify human references with respect to all automatic translations. Interestingly, the optimal set contains metrics working at all levels of granularity from 1-grams to 4-grams.
We use this metric set to compute the QUEEN measure for all systems. See results at the system level in Table 2. As expected, phrase-based systems attain best results, significantly better than the word-based system and 'SYSTRAN'.

[3]There are $2^{26} - 1$ possible combinations if we take into account all metrics.
[4]http://tc-star.itc.it/openlab2006/
[5]http://www.tc-star.org/
[6]http://www.europarl.eu.int/
[7]This system is an improved version of the 'PB' system which uses information at the shallow-parsing level to build better translation models as described by Giménez and Màrquez (2005).
[8]http://www.systransoft.com.

MT System	1-PER	1-WER	BLEU-3	GTM-2	MTR-exact	NIST-3	RG-L
WB	**0.34**	**0.58**	0.50	0.33	0.57	8.79	0.56
SYSTRAN	0.30	0.40	0.56	0.36	0.65	9.59	0.63
PB	0.26	0.36	**0.66**	**0.41**	0.69	10.66	0.66
PB++	0.26	0.37	**0.66**	**0.41**	**0.70**	**10.72**	**0.67**

Table 1: MT quality according to several metrics outside the IQMT Framework.

MT System	QUEEN
WB	0.31
SYSTRAN	0.39
PB	0.45
PB++	0.46

Table 2: MT quality according to the optimal metric set inside the IQMT Framework (QUEEN measure).

'PB++' slightly outperforms 'PB', although not very significantly. Interestingly, the 'SYSTRAN' systems performs significantly better than the 'WB' system. This means that, in this test set, translations produced by 'SYSTRAN' are more human-like than those produced by the word-based SMT system, even though 'SYSTRAN' is not designed for the specific domain.

Moreover, the QUEEN measure at the sentence level allows the user to perform a detailed error analysis by inspecting particular cases. Table 3 shows an interesting case of error analysis, in which all systems attain a QUEEN score under 0.2 except the 'PB++' system which scores 0.83. The QUEEN measure identifies the features which characterize human translations. QUEEN favours those automatic translations which share these features that are common to all references. In this case the 'PB++' system output is rewarded for providing exact translations, according to all references, for *'gestión de las crisis'* ('crisis management') and *'esperan señales'* ('they expect signs'). On the other hand, the automatic translations which do not share these common features are penalized.

Finally, the quality of the given test set of systems, references and metrics (JACK measure), considering the optimal metric set, is 0.77. This means that, in most cases (77%), system outputs are heterogeneously distributed closely around human references according to all metrics, and consequently, the test set is representative and reliable.

5. Ongoing work

Currently, we are devoting our main efforts to the development of syntax-based metrics. We are experimenting with metrics based on dependency trees provided by *MINIPAR* (Lin, 1998). We are also developing metrics based on shallow parsing annotation, i.e. part-of-speech, lemma and chunk information, provided by the *SVMTool*[9] (Giménez and Màrquez, 2004), the *Freeling*[10] package (Carreras et al., 2004) and the *Phreco* software, respectively. On the one hand we intend to capture morphosyntactic and syntactic similarities between automatic and reference translations. On the other hand we aim to capture partial aspects such as the proportion of correctly translated nouns and verbs, or the proportion of correctly translated noun and verb phrases. In the future, we intend to move on to other linguistic levels, such as Full Parsing or Semantic Role Labeling.

[9]The SVMTool may be freely downloaded at http://www.lsi.upc.es/~nlp/SVMTool/ .

[10]Freeling Suite of Language Analyzers may be freely downloaded at http://www.lsi.upc.es/~nlp/freeling/

Acknowledgements

This research has been funded by the Spanish Ministry of Science and Technology, projects ALIADO (TIC-2002-04447-C02) and R2D2 (TIC-2003-7180). The TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government. Authors are thankful to the TC-STAR Consortium for promoting the *'Openlab 2006'* Initiative which provides valuable data sets for both MT system development and evaluation.

6. References

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo. 2005. Qarla: a framework for the evaluation of automatic sumarization. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*, Michigan, June. Association for Computational Linguistics.

Bogdan Babych and Tony Hartley. 2004. Extending the bleu mt evaluation method with frequency weightings. In *Proceedings of 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 Machine Translation and/or Summarization*.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th LREC*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd Internation Conference on Human Language Technology*, pages 138–145.

C. Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Jesús Giménez and Lluís Màrquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of 4th LREC*.

Jesús Giménez and Lluís Màrquez. 2005. Combining linguistic data views for phrase-based smt. In *Proceedings of the Workshop on Building and Using Parallel Texts, ACL*.

source	los ciudadanos esperan de nosotros algo más que la simple **gestión de las crisis ; esperan señales** y una política sostenible en estos ámbitos .

systems

WB	the citizens expect of us something more than the simple **management of the crisis** and a sustainable policy in these areas . **expectantly signals**
SYSTRAN	the citizens wait for of us something more than the simple **management of the crises; they wait for signals** and a sustainable policy in these scopes.
PB	the citizens expect us any more than simply **managing crises ; they hope signals** and a sustainable policy in these areas .
PB++	the citizens expect us something more than simply **crisis management ; they expect signs** and a sustainable policy in these areas .

references

R0	the public expect more than just **crisis management ; they expect signs** , and a sustainable policy in these fields .
R1	citizens expect something more of us than just simple **crisis management ; they expect signs** and sustainable policies in these areas .
R2	the citizens expect from us something more than a simple **crisis management ; they expect signs** and a sustainable policy in these matters .

Table 3: A case of error analysis, according to the QUEEN measure, in which the 'PB++' system outperforms the rest.

Jesús Giménez, Enrique Amigó, and Chiori Hori. 2005. Machine translation evaluation inside qarla. In *Proceedings of the International Workshop on Spoken Language Technology (IWSLT'05)*.

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*.

G. Leusch, N. Ueffing, and H. Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*.

Chin-Yew Lin and Franz Josef Och. 2004a. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statics. In *Proceedings of ACL*.

Chin-Yew Lin and Franz Josef Och. 2004b. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of COLING*.

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 Machine Translation and/or Summarization*.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of HLT/NAACL*.

S. Nießen, F.J. Och, G. Leusch, and H. Ney. 2000. Evaluation tool for machine translation: Fast evaluation for mt research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2003. Final report of johns hopkins 2003 summer workshop on syntax for statistical machine translation. Technical report, Johns Hopkins University.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, ibm research report, rc22176. Technical report, IBM T.J. Watson Research Center.