A cheap MT evaluation method based on the notion of machine translationness

Joaquim Moré López
Linguistic Service
Open University of Catalonia (UOC)
jmore@uoc.edu

Salvador Climent Roca
Languages and Cultures Department
Open University of Catalonia (UOC)
scliment@uoc.edu

Abstract

In this paper, we present the notion of *machine translationness* (MTness) and its application in an MT evaluation method. Machine translationness is defined as the characteristics of texts generated by MT systems that are unlikely to be found in human-generated texts. We first show a typology of instances of MTness from an analysis of the output of a Catalan-Spanish MT system. Then we propose an inexpensive MT evaluation method which detects and counts instances of MTness by performing Internet searches. The goal is to obtain a sketch of the quality of the output, which, on occasions, is sufficient for the purpose of the evaluation. Moreover, this evaluation method can be adapted to detect drawbacks of the system in order to develop a new version, and can also be helpful for post-editing machine-translated documents.

1 Introduction

In this article we present the notion of *machine translationness* (MTness) and its application in MT evaluations. We coined the term *machine translationness* to refer to the linguistic characteristics of texts generated by MT systems that are unlikely to be found in human-generated texts. We will show some instances of machine translationness and will explain an MT evaluation method which is based on this notion.

Our method is intended to get a reliable snapshot as to the quality of an MT system by displaying a list of instances of MTness. The resources needed to detect these instances of MTness are not difficult or expensive to get. These are the bulk of web pages published in the target language and an Internet search engine.

Although the work is still at a preliminary stage, the results obtained make us believe that the method is both money-saving and suitable to perform fast, on-the-fly evaluations that provide a 'first impression' of the quality of the output. This first impression may be enough for the evaluation purposes of large organisations with MT (and consequently MT output evaluation) needs.

2 What is machine translationness?

We define machine translationness as the features in a text that indicate to the reader that the text is a translation and that this translation has not been performed by a human. In other words, MTness refers to the textual characteristics that would prevent a translation made by a machine from passing the Turing test. For instance, if a Catalan-Spanish MT system translates *moren de set* ('they die of thirst') as *mueran de siete* ('they die of seven') the system adds an example of MTness in its output.

3 Motivation

The idea of performing MT evaluations based on the notion of MTness arose from the evaluation needs of the MT system which operates at the Open University of Catalonia (UOC) in the Catalan-Spanish/Spanish-Catalan language pairs. The Open University of Catalonia is a virtual university which translates most of its educational material in Catalan into Spanish for students who are not Catalan speakers. Conversely, documents originally written in Spanish are translated into Catalan for the Virtual Campus in Catalan. The documentation is so immense that MT has been the solution to save time and money in terms of

the institution's translation needs. Since the costs of post-editing depend on the quality of the output, the UOC Language Service has been very keen to evaluate the quality of the MT system and, in order to save on correction costs, has worked on the detection of systematic errors that can be resolved automatically in a post edition module. Likewise, the Language Service also provides the system with new terminology and resources such as translation memories to improve the quality of the output.

We wanted to perform evaluations with a double-folded goal. On the one hand, to quickly assess the improvements made by the incorporation of new lexical items in the lexicon and the updating of the post edition module with new correction rules. This assessment must also be financially viable for the institution. On the other hand, to detect systematic errors that have a deep impact on the quality of the output. This detection should be automatic so that the post edition module can be updated as fast as possible.

4 Instances of MTness

In order to specify what is to be considered an instance of MTness, we analysed a corpus of translations performed by two Catalan-Spanish MT systems. These systems have different methodologies; one of them -Internostrum[1]- is based on the declaration of linguistic knowledge- and the other -N-II[2] by the TALP Group of the Universitat Politècnica of Catalunya- is a statistical MT system. The reason why we studied the output of both systems was to cope with as many MTness instances as possible, not only those that are peculiar of a particular methodology.

500 segments taken from newspapers, tourism web pages, administrative documents and economy reports in Catalan were translated by the two systems. The output was then analysed and their MTness instances were detected and classified manually.

The MTness instances were grouped into these classes:

- Untranslated words
- Typo errors
- Contextually odd translations
- Errors in linguistic competence

Here we comment some of them

Untranslated words and typo errors

Besides the words in the original that appear untranslated, we also considered the apparition of strange characters, upper-case and lower-case inconsistencies, etc. as examples of MTness.

Contextually odd translations

This covers illegitimate word-for-word translations such as the mistranslation of acronyms (e.g. translation of *memòria RAM* as *la memoria RAMO* which literally means 'bouquet memory'), translations of idioms (e.g.: translation of *fer el préssec* which means "make a fool of oneself" as *hacer el melocotón*, literally 'to make the peach') and translations of expressions that are not consistent with the communicative intention of the speaker (e.g. translation of *no té cap mena d'importància* ('it's nothing') as *no tiene el menor asomo de importancia* instead of *no tiene ninguna importancia*).

Many translations sound bizarre because of a homonym confusion of a word in the original. For instance, the Catalan noun *vol* (flight) coincides with the third person singular of the verb *voler* (want) in the present tense. Thus, *sortida vol* is translated as *sortida quiere* (departure wants). Another example is the translation of the Catalan sentence *morin de set* (they die of thirst) as *mueran de siete* (literally, they die of seven). The system has wrongly mistaken 'set' (thirst) as the numeral (seven).

The *ser/estar* confusion is also included in this class. *És* (is) can be translated both as *es* or *está*, i.e. the permanent vs. temporary 'to be'. The system often takes the wrong option as in *el disco es lleno* (the disk is full) instead of *el disco está lleno*.

We found translations with words that do not correspond to any word in the original. These extra words make the translation confusing. Conversely, there are unintelligible translations because a word in the original does not appear in the translation. For instance, *un magnífico del arquitectura islámica* (a magnificent of the Islamic architecture). These two kinds of errors are typical of the statistical MT system and probably they are due to its performance based on what it learned from a training corpus.

We also saw examples like this one

[1] www.internostrum.com
[2] http://www.n-ii.org/

(1) *Setecientos bocados empiezan a patrullar en once zonas de Barcelona.* (seven hundred bites start to patrol round eleven areas in Barcelona).

According to the selectional restrictions of the verb *patrullar* (to patrol) the subject of the sentence must be human. So *setecientos bocados* violates this restriction. The original for *setecientos bocados* is *set-cents mossos*. *Mossos* is the name for the policemen that work for the autonomous police in Catalonia. So the translation of the subject as 'seven hundred bites' makes the sentence unintelligible.

Finally, we detected anaphoric errors (e.g. *yo* ***los*** *estaría muy agradecido* instead of *yo* ***les*** *estaría muy agradecido* ('I would be grateful to them')), and cases of wrong compound words. For instance, *archivos excielo* instead of *archivos excel* (excel files). *Ex* is taken as a prefix for *cel* (sky).

Errors in linguistic competence

This covers wrong number, gender and person agreement (e.g.: *un paseo por* ***los animadas calles*** instead of *un paseo por las animadas calles*); no apocopation as in the wrong use of *grande* and *primero* instead of *gran* and *primer* in *un grande momento* (a great moment) and *el primero ministro* (the Prime Minister). Errors such as *de el* instead of *del* (of the) or *y hizo* (and he/she did) instead of *e hizo*.

5 Evaluation method design

The goal of the evaluation method is to get a list of instances of MTness of both system A and system B (or version S1 and S2 of the same system) and compare the number of instances of A and B (or the number of instances of S1 and S2). The system or version with the lesser number of instances of MTness the better.

In order to find instances of MTness, we relate the probability of the generation of a piece of MT output by a fluent speaker with the number of occurrences in a representative corpus of the target language. The instances of MTness we deal with are translation solutions chosen by the system among a set of possible translations. So we take into account the audience's expectancy to find the MT translation solution in a fluent text. This expectancy is inferred by comparing the number of occurrences of each possible translation solution in the representative corpus.

Thus, we have focused on instances of MTness that comply with this condition: given a source chunk SC and a chunk TC_i which is the translation of SC generated by an MT system out of TC_1, TC_2,...TC_n possible translations, TC_i is an instance of MTness when the number of occurrences of TC_i in a representative corpus of the target language is overwhelmed by the number of occurrences for any of the other possible solutions. In case there are no other possible translation solutions and the number of occurrences of TC_i is less than five, then TC_i is considered a candidate to be an instance of MTness. This is due to the lack of evidence for TC_i to be considered a real instance of MTness (see section 7).

For practical reasons, we have taken all the web pages published in the target language as the representative corpus. So the number of occurrences of a chunk in this representative corpus can be inferred from the number of web pages containing it according to a search engine, provided that the target language is widely present on the World Wide Web. No results or less than 5 results means that the appearance of an MT chunk in a fluent target language text is highly improbable, so it may be an instance of MTness; whereas a chunk with more than, say, 1,000 results is not considered as such. For example, the chunk *quedó en no nada* is not found on any web page when using the Yahoo and Google search engines[3].

The method has the following stages: MT output tagging, creation of MT output chunks, alternative chunk creation, MTness detection and, when comparing different systems or versions of the same system, results comparison.

MT output tagging. The MT output is syntactically tagged by an automatic tagger. We used the open source language tool FreeLing (Atserias, Casas, Comelles, González, Padró and Padró, 2006) for the evaluation prototype (see section 6).

Creation of MT output chunks. The tagged MT output is split into MT chunks. The chunks established so far are the following: noun phrases, verbs (simple and complex), adjectival phrases with the role of verbal complement, adverbial phrases, and adjunct prepositional phrases. Other chunks are strings where two chunks of the type described coexist with no

[3] The results that appear in this article are from consultations dated on 10/02/06.

punctuation mark in between them and express a relation between two concepts. So far we have taken into consideration the coexistence of a noun phrase with a verb, a noun phrase with a verb and an adjectival phrase, two or more noun phrases together and finally a verb with a prepositional phrase as its argument.

Alternative chunk creation. For each MT chunk, alternative translations are created. An alternative for a chunk *C* is a new chunk *C'* created automatically by one of the following actions, which, henceforth, will be called A1 and A2:

A1. Substitute a translated uppercase word for its corresponding source word (e.g.: Catalan: *memòria RAM* ('RAM memory'); Spanish C: *memoria RAMO*; Spanish C': *memoria RAM*).

A2. If there is a word TW, whose corresponding source word SW has a different translation, TW', substitute TW for TW'.

(2) **Catalan:** Sortida vol
(Departure flight)
Spanish C: Salida quiere
(Departure wants)
SW: vol ('flight')
TW: quiere ('wants')
TW': vuelo ('flight')
Spanish C': Salida vuelo

So far we have outlined these two actions, but other actions could be performed to cope with phenomena that go beyond lexical selection and affect syntax. For instance, the action of adding a definite article before a determinerless noun in the original (e.g., *problems with teenage behaviour -> problemas con* ***el*** *comportamiento adolescente*) or putting adverbials in a new order (*llevar más mucho tiempo -> llevar mucho más tiempo*).

In order to create alternative chunks automatically the following resources are needed: a source and target language wordlist, with the form, lemma and POS tags for each word, and a list of <source word form, target word form> pairs, where 'target word form' is the translated equivalent of the source word form. For instance, the alternative *mueran de sed* for *mueran de siete* is created when the following pairs <set, siete> and <set, sed> are found.

Detection of instances of MTness. In a way similar to the selection of MT translation candidates (Greffenstette, 1999), for each new MT chunk, the detector obtains the number of web pages that contain it. This information is provided by an Internet search engine. When the MT chunk has alternatives, they are also searched for by the engine and their results are compared to the results of the MT chunk. If the number of results for an alternative chunk overwhelms the number of results for the MT chunk, the latter is considered an instance of MTness. The instances of MTness are stored in a list. In case the MT chunk has no alternatives and the number of results is less than 5, the MT chunk is stored in a list of candidates to be instances of MTness.

Results comparison. The number of instances of MTness for system A or the latest version is compared to the number of instances for system B or the previous version. The fewer the number, the better the system or version. The lists of candidates to be instances of MTness for A and B are also compared. If one of the lists has a candidate which is not in the other list, this candidate is counted as a real instance of MTness.

6 Evaluation method prototype

In order to test the feasibility of the method, we tried to find instances of MTness in the Spanish translations for 500 Catalan segments performed by a particular MT system. We chose the open-source system Apertium[4], as this system's resources can be obtained freely; thus, the Catalan and Spanish word forms and the list of <source word form, target word form> pairs could be generated automatically. From the 396 errors detected manually we focused on the following types of error.

- Illegitimate word-for-word translations
- Homonym confusion
- No apocopation

These phenomena caused 61.2% of the errors detected. The rest spanned a range of errors that could be easily detected by a word and grammar corrector, such as untranslated words due to typographical errors in the originals (19.2%) and untranslated words due to their not appearing in the bilingual dictionary (10%), non-agreement in gender or verbal person (4.3%), and contraction and syntactic phonology errors such as *de el* instead of *del* (of the) or *y hizo* (and he/she did) instead of *e hizo* (0.7%).

This amounts to more than 95% of the errors. Among the rest, there are violations of selec-

[4] http://apertium.sourceforge.net/

tional restrictions, and anaphoric errors. Table 1 shows some instances of MTness detected by the method. The correct translation for most of these instances has been found by selecting the alternative that overwhelms the MT chunk with more results.

Error Typology	Source Chunk	MT Chunk	MT Chunk Results	Alternative Chunk	Alternative Chunk Results
Misinterpretation of the sense of a source word	Morin de set (die of thirst)	Mueran de siete	0	Mueran de sed	164
	Jornada sagnant (bloody day)	Jornada sangrante	0	Jornada sangrienta	32,100
Word form confusion	Sortida vol (departure flight)	Salida quiere	61	Salida vuelo	310
	Sortir a sopar (go out for dinner)	Salir a cena	7	Salir a cenar	19,200
	endeutament net (net debt)	endeudamiento limpio	0	endeudamiento neto	1450
No apocopation	Una gran festa (a big party)	Una grande fiesta	167	Una gran fiesta	188,000
	Primer contacte (first contact)	Primero contacto	416	Primer contacto	492,000
Illegitimate word-for-word translation	Fer el préssec (make a fool of oneself)	Hacer el melocotón	0		
	Memòria RAM (RAM memory)	Memoria RAMO	6	Memoria RAM	1,320,000
Improper use of ser/estar	El disc és ple (the disk is full)	El disco es lleno	0	El disco está lleno	398
	És previst d'arribar (it is expected to arrive)	Es previsto llegar	0	Está previsto llegar	200

Table 1. Instances of machine translationness detected via web searches

7 Discussion

Among the current trends in MT evaluation methods, our approach is not based on the Assumption of Reference Proximity (ARP) such as BLEU (Papineni et al. 2001), NIST (Doddington, 2002), WER (Nießen et al. 2000), GTM (Melamed et al., 2003), ROUGE (Lin and Och, 2004) and METEOR (Banerjee and Lavie, 2005). Rather, we follow the trend based on the Human Likeness criterium (Amigó et al. 2006): a machine translation that could have been generated by a human is a good translation whereas a machine translation which cannot be attributable to a human is a bad translation.

One of the advantages of evaluations based on the Human Likeness criterium is, in our opinion, that they are more significant and reliable about the translation quality of the output than those based on the PRA. During a BLEU-based evaluation of an MT system performed by the Linguistic Service at UOC, we realised that the reference translations had to be revised (Moré and Climent, 2006). The reason was that a significant number of acceptable MT translations would have been unfairly penalised because of the presence of dubious references. This revision made our evaluation time-consuming and expensive. Apart from this, ARP metrics do not convey information about systematic errors and their impact on the translation quality. Besides, Giménez and Amigó (2006) argue that these metrics do not take into account any information at linguistic levels further than lexical.

Another advantage of evaluations based on the Human-Likeness criterium is that there is no need to gather a large corpus to determine whether the evaluator is assessing a machine or a human translation (Reeder, 2001) and the fact that it leads to the detection of systematic translation errors that can be used in automatic correction modules to reduce post-editing costs (Gamon, Aue, and Smets, 2005).

However, if the classification is performed automatically by a classifier that has learned previously the characteristic features of machine and human translations (Corston-Oliver, Gamon, and Brockett, 2001; Kulesza and Shieber, 2004), we face the time-and-money consuming task of compiling training corpora, linguistically and semantically annotated, with huge numbers of instances of machine and human translations (Gamon et al., 2005).

Another approach to automatically identify human translations from automatic ones is by setting a similarity metric that identifies and uses the features which are common to all human references (Giménez and Amigó, 2006). This method is also expensive for us as it requires at least three human references. Thus we tried to design an alternative method that automatically identified mistranslated sentences, provided a list of systematic errors and offered a fast diagnosis of the system's behaviour that saved time and money, with costless resources.

Our method if combined with a spelling and grammatical corrector can detect over 90% of the translation errors from our evaluation test and, correspondingly, most of the instances of MTness. The detection is carried out with free

resources (web pages on the World Wide Web, wordlists and a free, open-source tagger) and correction tools that are largely widespread for editing documents. Likewise, the detection of instances of MTness also provides information that can be useful for developing an automatic post-editing module and to set a strategy to improve the output of the system. The 'instance of MTness/alternative with most results' pair could be presented to proofreaders of machine-translated documents who could accept or decline the alternative. The accepted alternatives would be propagated throughout the document and stored in a repository in order to perform automatic correction of machine-translated documents. Thus, the costs are greatly exceeded by the benefits of the results obtained and the possibility of reusing them. This is why we present the method as being cheap.

Furthermore, the results of the evaluation are significant because the method is consistent with the idea that human evaluators detect aspects that characterise machine translations and that they penalise translations with a high probability of belonging to the machine class rather than the human class. However, we are aware that this method is intended to perform fast, on-the-fly evaluations in order to get a reasonable 'first impression' of the quality of the output, which, on occasions, is sufficient for the purpose of the evaluation and, on other occasions, is simply the first stage for a sounder analysis of the output, if purposes so require.

Nonetheless, there are two aspects that deserve special attention. These aspects concern the possible distortion of the results due to errors made by the automatic tagger and the presence of grammatical errors and other problematic features on web pages.

As for errors committed by the tagger, it is not absolutely necessary to label all the chunks with their proper syntactic label. The tagger merely establishes a criterion to split the sentences into chunks that will be turned into queries for the search engine. The important thing is for the query to contain a semantically significant word (noun, verb, adverb, and so on) together with the words that the tagger considers as its semantic complements regardless of whether the label is absolutely correct or not. So, if a word is tagged, say, as a noun when it is actually a verb, it does not make a difference for our purposes if nominal complements are taken as verbal complements instead; in other words, if a semantic relationship between them is detected. For example, it makes no difference if *sortida vol* is tagged as a noun phrase followed by a verb, or if it is simply labelled as a noun phrase. The evaluator will trigger the same query.

Secondly, as regards web pages, the mere appearance of a certain chunk is not always significant in determining its MTness or its non-MTness. For example, the Spanish mistranslation of *pla d'estudis* ('study plan' in Catalan) as *plano de estudio* is found on the Internet because it coincides with the Portuguese term. Likewise, we have to take into account the presence of blogs, web pages with a careless use of language and even machine translated web pages which have not been post-edited. For example, *disco llevar* ('disk take') as a translation of *disc dur* ('hard disk') appears in a machine-translated web page. However, most of these chunks are overwhelmed by the number of examples of the correct translation alternative (e.g.: disco llevar 63; disco duro 8,540,000) or do not appear when the chunk coexists with another chunk in a larger query. An example of the latter is the Spanish translation of the Catalan *el nou* ('the new') as *lo nueve* ('the nine') because nou can be interpreted as the numeral 'nine' or the adjective 'new'. *Lo nueve* has 369 results, but *lo nueve gobierno* ('the nine government') has no results. However, we would wish to stress that although we have presented the Web as the largest representative corpus, we are not saying that other kinds of corpus cannot be representative of language use depending on the evaluation needs. If the corpus came from published documents in only one language, so they underwent a post-editing process, the problems we have just mentioned would not arise.

Nonetheless, the lack of results in a representative corpus is not always a direct indication of an instance of MTness. For example, a perfectly grammatical Spanish chunk like *mataron a Rigobert Mallafré* ('they killed Rigobert Mallafré') returns no results because Rigobert Mallafré is an individual not referred to on any web page. This is the reason why this chunk is not considered a real instance of MTness but just a candidate to be one. We are considering to tackle with these cases by performing semantic tests. These tests would consist in replacing proper nouns or common nouns by a constituent semantically equivalent according to an online lexical database like Wordnet, and assessing the new query. So, by hypothesing that *Rigobert Mallafré* is a human being, we could substitute *Rigobert Mallafré* for *un policía* (a policeman) and we would

get 12,700 results. That would tell us that the chunk is not an instance of MTness.

Similar tests could be performed for perfectly grammatical Spanish chunks containing a reference to an entity present on the Web but with no results (e.g. *el patrimonio de la Provenza* (Provence's cultural heritage)*:* 0 results but *el patrimonio de la Rioja*: 552 results).

Although the prototype we have presented here mostly deals with instances of MTness at the lexical level, we aim to detect wrong translation solutions that affect syntax. If we could obtain the semantic restrictions of a verb onto its arguments from an online lexical resource, we would be able to detect violations of selectional restrictions that affect syntactic relationships (subject-verb, verb-object) .

8 Conclusions and future work

The evaluation method presented is still in a preliminary phase, but the initial results obtained are encouraging enough to keep on working on its full development. Contrary to other MT-evaluation proposals that do not use human translation references and which are based on the ability of a classifier to distinguish machine translation qualities that are not characteristic of humans, our method does not need large corpora of human and machine translations to train a classifier. The resources and the performance of the method are inexpensive and provide a quick assessment of the quality of the output that may be sufficient depending on the purpose of the evaluation. Thus, it is reasonable to expect that an evaluation offering valuable results, without requiring great amounts of time or money, is possible.

Apart from the economic advantages, the data obtained by applying this method can be reused for other purposes. The list of instances of MTness provides information about the drawbacks of an MT system and is very useful for developers in improving their performance (micro-evaluation). Likewise, the method could be adapted to test the quality of language in pages published on the Web. For instance, by detecting instances of MTness, we can find evidence of web pages that have been translated automatically and not post-edited.

As for future work, first we will refine the method so that chunks like *mataron a Rigobert Mallafré* and *el patrimonio de la Provenza* are not considered candidates to be instances of MTness. Then we will carry out a full evaluation of the method proposed in the language pair already studied and in other language pairs. We will also try to detect more instances of MTness that go beyond lexical errors and affect syntax.

Finally, we intend to reuse the information obtained from all these error-detection strategies to perform semi-automatic post-edition tasks in order to save time and money in corrections.

9 Acknowledgements

This work is carried out under the project RESTAD (Resources for computer-based translation applied to teaching) funded by the Ministry of Education and Universities of the Generalitat de Catalunya and developed by the Universitat Autònoma of Barcelona, the Universitat of Girona, the Universitat Oberta of Catalunya and the Universitat Politècnica of Catalunya.

References

E. Amigó, J. Gimenez, J. Gonzalo, and L. Márquez. 2006. MT-Evaluation: Human-like vs. human acceptable," Proceedins of ACL.

J. Atserias, B. Casas, E. Comelles, M. González, L. Padró and M. Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the 5th International Conference on Language Resources and Evaluation.* Genoa, Italy.

S. Banerjee and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlations with Human Judgments. In *Proceedings of Workshop on Intrinsicand Extrinsic Evaluation Measures for MT and/or Summarization* at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005). Ann Arbor, Michigan

S. Corston-Oliver, M. Gamon and C. Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. *Proceedings of the Association for Computational Linguistics.* Toulouse, France.

G. Doddington. 2002. Automatic evaluation of language translation using n-gram cooccurrence statistics. *LREC-2002: Third International Conference on Language Resources and Evaluation.* Workshop: Machine translation evaluation: human evaluators meet automated metrics, Las Palmas Canary Islands, 27 May 2002; 9pp.

M. Gamon, A. Aue and Smets, M. 2005. Sentence-level MT Evaluation Without Reference Translations: Beyond Language Modelling. *Proceedings of the 10th Annual EAMT Conference.* Budapest.

J. Giménez and E. Amigó. 2006. IQMT: a framework for automatic machine translation evaluation. *LREC-2006: Fifth International Conference on Language Resources and Evaluation*. Proceedings, Genoa, Italy.

G. Grefenstette. 1999. The WWW as a Resource for Example-Based MT Tasks. *Proc. Of Aslib Conference on Translating and the Computer.* London.

A. Kulesza and S. M. Shieber. 2004. A learning Approach to Improving Sentence-Level MT Evaluation. Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation. Baltimore.

C. Lin and F. J. Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In Proceedings of COLING.

I. D. Melamed, R. Green, and J. P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of HLT/NAACL*.

J. Moré and S. Climent. 2006. A cheap MT-evaluation method based on Internet searches. *EAMT-2006: 11th Annual Conference of the European Association for Machine Translation,* Oslo, Norway.

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*.

K. Papineni, S. Roukos, T. Ward and W-J. Zhu. 2001. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA.*

F. Reeder. 2001. In One Hundred Words or Less. *MT Evaluation Workshop MT Summit VIII*. Santiago de Compostela.