A Cheap MT-Evaluation Method Based on Internet Searches

Joaquim Moré and Salvador Climent

Open University of Catalonia
jmore@uoc.edu scliment@uoc.edu

Abstract. In this paper, we first argue that human translation references used to calculate MT evaluation scores such as BLEU need to be revised. This revision is time and resource-consuming, so we propose, instead, using an inexpensive MT evaluation method which detects and counts examples of characteristic MT output, referred to herein as *instances of machine-translationness*, 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

Depending on the purpose of an MT evaluation, getting a 'first impression' of the quality of the output may be enough. Fast, on-the-fly evaluations can provide concrete proof of the quality of a system's translation and save both time and money. As a result of the actual application of MT and MT evaluation methods in a large organisation, the Open University of Catalonia (UOC), we can present in this paper a method that allows for the performing of fast, large-scale on-the-fly evaluations without using Human Translation References (HTR) or large corpora of machine translated and human translated texts.

The method is based on the detection of instances of so-called *machine translationness* in the output sentences. We have coined the term *machine translationness* in order to refer to the character certain MT outputs have which are unlikely to be considered as human translations. Instances of machine translationness can be detected by using Internet search engines.

After presenting the reasons and discussing the method, we present a prototype which focuses on five types of common MT errors. According to the results obtained by this prototype we can conclude that, although the work is still at a preliminary stage and cannot yet be fully assessed, it is very promising as a time and money-saving methodology for an organisation with MT (and consequently MT output evaluation) needs.

2. MT evaluation needs at the UOC

The Open University of Catalonia (UOC) 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. Likewise, the Language Service also provides the system with new terminology and resources such as translation memories to improve the quality of the output, so it is necessary to perform continuous evaluations in order to assess the improvements made.

When the UOC Language Service undertook the evaluation of the MT system, the method chosen was the calculation of BLEU (Papineni, Roukos, Ward and Zhu, 2001), given the reduction in time and money compared to human evaluation. The Language Service also prepared the resources necessary to perform future evaluations of MT systems for other language pairs such as Catalan-English and English-Catalan.

For each source language (English, Spanish and Catalan), a set of 500 segments taken from newspapers, tourism web pages, administrative documents and economy reports was prepared. The Language Service also took care of the reference translations for each segment in the following language pairs: Catalan-Spanish, Spanish-Catalan, English-Catalan and Catalan-English. These translations were performed by four professional translators, who were native speakers of the target language and had a wealth of experience, with degrees and diplomas accrediting their translation expertise.

The segments corresponded roughly to a sentence and were sorted in a random order. So human translators were put in the situation of MT systems that translate sentence by sentence, without bearing in mind what comes before or after. However, although the segments were decontextualised, they were not meaningless. From the thousands of segments obtained we selected 500 that were meaningful and which could be translated faithfully in terms of the original.

We analysed the references provided by the professional translators to guarantee that the references would not distort the evaluation because of one of the following reasons:

- The reference is as illegitimate as the MT hypothesis; in this case, the lack of coincidences with any reference may penalise a correct hypothesis.
- The translators express how they have interpreted the source segment by using words and constructions that do not correspond to a word-for-word translation, even when a word-for-word translation would be legitimate. The probability of producing unreliable references because the translator misunderstood the decontextualised original segment is quite high. Besides, it is less likely for a legitimate word-for-word machine translation to match a reference.

In a revision on the fly, we concluded that all 8,000 segments (2,000 for each language pair direction) had to be revised because there were a significant number of references that could distort an MT evaluation due to one of the aforementioned reasons. Examples were found from nearly all the translators working in the same direction, and also among translators that worked in different directions. The examples we present each belong to a different translator and come from two different language pair directions (Spanish-Catalan, English-Catalan).

(1) ***London upset Paris on Wednesday for the right to host the 2012 Summer Olympics*** (Original in English)

El dimecres, Londres va preocupar a París pel dret a acollir els jocs olímpics d'estiu el 2012 (Reference in Catalan)

In the Catalan reference, *upset* is translated as *va preocupar* which means *worried*, but the translator should have used a Catalan word or expression meaning a different sense of the verb *upset*, i.e. *defeat suddenly and unexpectedly*. In this case, the human translator, as MT systems often do when translating a decontextualised segment, did not use the proper sense of the original word.

Here are two further examples, (2) and (3), which are problematic because of the translator's decision not to perform a word-for-word translation:

(2) ***PRICE, CHRISTINE, 81, went to be at rest on August 29, 2005.*** (Original in English)

PRICE, CHRISTINE, de 81 anys, va ser enterrada el 29 d'agost de 2005. *(*Reference in Catalan*)*

(3) ***Magnífico hotel ecológico rodeado de exuberante naturaleza.*** (Original in Spanish)

Magnífic hotel ecològic envoltat de vegetació exuberant. (Reference in Catalan)

Magnífic hotel ecològic envoltat d'exuberant naturalesa. (MT Catalan hypothesis)

Magnificent ecological hotel surrounded of exuberant nature. (English word-for-word translation)

In (2), *went to be at rest* is translated as *va ser enterrada* (was buried). The segment was taken from an obituary, so it was impossible for the dead person to have been buried yet; the translator should have used the Spanish verb *morir* (to die) or a synonym. As for (3), the proofreaders spent a long time discussing whether the translation of *naturaleza* (nature) as *vegetació* (vegetation) was legitimate or not. This is an example of how deciding about the legitimacy of a reference may take longer than judging the machine translation hypothesis as correct. Besides, in this case if *naturalesa* did not appear in the references, a system that performed a correct translation would be unfairly penalised.

We concluded then that the revision of the HTR required, and the effort taken in discerning their legitimacy, made evaluation with human translation references time-consuming and expensive. Thus, we tried to design an alternative method without HTR. A method that automatically identified mistranslated sentences and offered a faster diagnosis of the system's behaviour to save time and money.

3. Evaluation method design

Among the proposals for automatic evaluations without reference translations, we prefer those where the assumption is as follows: if a translation is identified as produced by an MT system, not by a human translator, then it is a bad translation. So the evaluation consists in classifying the translations in the evaluation set as human or machine translations (Carston-Oliver, Gamon, and Brockett, 2001), (Kulesza and Shieber, 2004). The more confident the evaluator is in classifying a translation as produced by a machine, the worse is its quality; and conversely, the more confident the evaluator is in classifying it as human the better it is. We find this approach compelling because it is based on a common-sense assumption and although human-like machine translations may fail in semantic fidelity to the original, we consider it a good way of getting a reliable snapshot on the fly as to the quality of the output generated. As we are interested in performing continuous evaluations, this snapshot is sufficient for our needs; leaving the human testers to evaluate the output in greater depth when, for example, we are interested in knowing the fluency and fidelity of pieces of texts that are recurrent in the institution's documents, especially in those cases where the content is particularly sensitive or important. Other advantages of this approach include the fact 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). Lastly, these evaluators classify translations after having learned the characteristic features needed to perform the classification. So once the evaluator has been trained, regular evaluations can be carried out quickly and at a low cost. However, machine learning of the characteristic features of machine and human translations requires training corpora with huge numbers of instances of both types, which are very expensive to compile and which require annotation with linguistic and semantic features, etc. (Gamon et al., 2005).

We propose an evaluation based on a list of instances of characteristic MT output retrieved without a training corpus. We call these *instances of machine translationness*. We have coined the term *machine translationness* (henceforth MTness) to refer to the quality of MT output unlikely to be generated by a fluent speaker of the target language because the system is unable to be critical about its own output and does not foresee, when faced with to

two or more possible translation solutions, the impact of choosing one of them – whether the audience would take its output as intelligible and well expressed or, on the contrary, hardly intelligible or even nonsense. This critic capacity distinguishes human and machine translators, the former being constant evaluators of their output whilst producing it and always hypothesising as to the reaction of the audience in their decisions. For example, *mueran de siete* (they die of seven) and *salida quiere* (departure wants), which are the Catalan-Spanish MT translations of *morin de set* (they die of thirst) and *sortida vol* (departure flight) respectively, are instances of MTness because their generation by a Spanish native speaker is highly unlikely and, likewise, because they show the inability of the system to foresee the reaction of the audience in terms of the decision to translate *set* as *seven*, and *vol* as *wants*, decisions human translators would not make, because they consider them absurd translations and know that the audience would consider them likewise.

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. Similarly, we relate the reaction of the audience to a translation solution with the audience's expecting to find it 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 zero or this number is overwhelmed by the number of occurrences for any of the other possible solutions

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 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 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 such. For example, the chunk *vuelan esconder* (fly hide) is not found on any web page when using the Yahoo and Google search engines (last consultation 10/02/06).

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 analysis tool FreeLing (Atserias, Casas, Comelles, González, Padró and Padró, 2006) for the evaluation prototype (see section 4).

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 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 (eg: 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'.

(4) ***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***).

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, target word> pairs, where 'target word' is the translated equivalent of the source word. For instance, the alternative *morir de sed* for *morir 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. If there are no results, the chunk is put in a list of candidates to be instances of MTness. 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 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.

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.

4. Evaluation method prototype

In order to test the feasibility of the method, we tried to find instances of MTness in the MT Spanish translations for the 500 Catalan segments prepared by the UOC Language Service (see Section 2). The translations were performed by the open-source system Internostrum[1], as this system's resources can be obtained freely; thus, the Catalan and Spanish wordlists and the list of <source word, target word> pairs could be generated automatically. We chose the Catalan-Spanish direction because these languages are very closely related and, consequently, the instances of MTness would stand out more obviously. From the 396 errors detected manually we focused on the following types of error.

- ***Misinterpretation of the lemma of a source word (34.4%)***

  Among the various senses of a source-language word, the system interprets the wrong one. When the Catalan sentence *morin de set* (they die of thirst) is translated as *mueran de siete* (literally, 'they die of seven'), the system has wrongly interpreted *set* as the number.

- ***Word form confusion (13%)***

  In the lexical selection of a target word, the system is misled by the coincidence in form of the source word with another source word whose meaning does not fit the context. 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* in (4).

- ***Illegitimate word-for-word translation (11.4%)***

This covers mistranslation of acronyms (eg: translation of *memòria RAM* as *la memoria RAMO* which literally means 'bouquet memory'), translation of idioms (eg: translation of *fer el préssec* which means 'make a fool of

[1] www.internostrum.com

oneself' as *hacer el melocotón*, literally 'to make the peach'), non-dropped prepositions in verbal complements (eg: *pensó en dimitir*, where the preposition *en* comes from the original *va pensar a dimitir*, which means 'he/she pondered resigning'), articles before proper nouns (*el Irán*), etc.

- ***No apocopation (1.7%)***

For instance, wrong use of *grande* instead of *gran* as in *un grande momento* (a great moment) or *primero* instead of *primer* as in *el primero ministro* (the Prime Minister).

- ***Improper use of the verbs 'ser' and 'estar' (0.7%)***

És (is) can be translated both as 'es' or 'está', ie, 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*.

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%). Lastly, 4.3% of the errors were varied and unsystematic 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.

5. Discussion

As we have seen in the previous section, 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 a semi-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.

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

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, 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 (eg: *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. For example, a translation need for the UOC is the translation of exam questions, originally written in Catalan, into Spanish, given that the exams have to be the same for students that speak Catalan and those that don't. The corpus could be all the exam questions for all the subjects taught at the university. Likewise, if the corpus came from published documents, which implies that 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 a machine translation error. 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. We are considering substituting proper nouns with an NP with a very frequent nominal head, even a proper noun, to represent 'person' or 'institution', etc., and assessing the new query. In other words, we could substitute *Rigobert Mallafré* for *un policía (*'a policeman') and we would get 552 results that tell us that the chunk is not an instance of MTness. If we could obtain the semantic selectional restrictions for verbs from an online lexical resource the proper noun could be substituted directly by an NP containing a noun head with the sense selected by the verb.

6. 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.

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. In order to do this, we are thinking of comparing the results in the translation chunk with the results of different ways of expressing its semantic content which are found by performing a non-exact matching Internet search (Moré et al., 2004). For example, the chunk *Ucrania primer ministro* as the Spanish MT translation of *Ukraine Prime Minister* has 72 results. By performing the non-exact matching search 'Ucrania primer ministro', the snippet of the first result on the results page is *Ucrania.- Yushchenko jura hoy su cargo como* ***primer ministro*** *de* ***Ucrania*** (Yuschenko sworn in today as Ukrainian Prime Minister'). *Primer ministro de Ucrania* is a prepositional phrase with 517 results; so the MTness of *Ucrania primer ministro* can be proven.

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.

7. References

Atserias, J., Casas, B., Comelles, E., González, M., Padró L., & Padró, M. (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.

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

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

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

Kulesza, A. & Shieber, S. M. (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.

Moré, J., Climent, S. & Oliver, A. (2004). A Grammar and Style Checker Based on Internet Searches. *Proceedings of the LREC2004*. Lisbon.

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

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