

-A cheap MT-Evaluation Method based on Internet searches

Joaquim Moré López

Open University of Catalonia (UOC)
jmore@uoc.edu

Salvador Climent Roca

Open University of Catalonia (UOC)
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 a cheap MT evaluation method which detects and counts characteristic MT output, referred here as *instances of machine-translationness*, by performing Internet searches. 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

As a result of a real application of MT and MT evaluation methods in a large organisation such as the Open University of Catalonia (UOC), we present in this paper a solution for evaluating MT output without using Human Translation References (HTR) nor 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 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 motivation 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 too preliminary to 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 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 bulk of documentation is so immense that MT has been the solution to save costs in time and money with the translation needs of the institution. Since the costs of postediting depends on the quality of the output, the UOC Linguistic Service has been highly concerned in evaluating the quality of the MT system and, in order to save correction costs, has worked on the detection of systematic errors that can be solved automatically. Besides, the Linguistic 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 Linguistic Service undertook the evaluation of the MT system, the method chosen was the calculation of BLEU (Papineni et al. 2001) because of its cheaper cost in money and time compared to human evaluation. The Linguistic 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 Linguistic 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 in the target language and had a large 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 and what comes after. However, although the segments were decontextualised they were not meaningless. Among the thousands of segments obtained we selected 500 segments that were meaningful and could be translated faithfully to the original.

We analysed the references delivered 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-to-word translation, even when a word-to-word translation would be legitimate. The probability of performing non reliable references because the translator misunderstood the decontextualised original segment is rather high. Besides, it is less probable for a legitimate word-to-word machine translation to match a reference.

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

First, we show the headline of a piece of news as an example of how human translators may behave the same way as MT systems when translating decontextualised sentences.

- (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.

Let's see two examples- (2) and (3)- which are problematic because of the translator's decision not to perform a word to 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-to-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 be buried yet; the translator would have used the Spanish verb *morir* (to die) or a synonym. As for (3), the revisors 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 necessary revisions of HTR, and the effort taken in discerning their legitimacy, made evaluation with human translation references more time-consuming and expensive. So we tried to design an alternative method with no HTR, identifies automatically badly-translated sentences and allows getting a fast diagnosis of the behaviour of the system which saves time and money.

3. Evaluation method design

Among the proposals of automatic evaluations without reference translations, we prefer those whose assumption is the following: 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 et al (2001), Kulesza and Shieber (2004)): the more confident is the evaluator in classifying a translation as produced by a machine, the worse is its quality; and conversely the more confident is the evaluator in classifying it as human the better it is. This approach is attractive for us 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 is a good way of getting a fiable snapshot on the fly about the quality of the output generated. As we are interested in performing continuous evaluations, this snapshot is enough for us, leaving for human testers evaluate the output in a deeper level when, for instance, we are interested in knowing the fluency and fidelity of pieces of texts that are recurrent in the documents of the institution and whose content is particularly outstanding. Other advantages of this approach are there is no need to gather a large corpus to determine whether the evaluator is facing a machine or a human translation (Reeder, 2001) and implies detecting systematic translation errors that can feed an automatic correction module that can save post-edition costs (Gamon et al., 2005). Lastly, these evaluators classify translations after having learned the characteristic features that will lead them to perform the classification. So once the evaluator has been trained, regular evaluations can be carried out quickly and with very low cost. However, machine learning of the characteristic features of machine and human translations requires training corpora with huge instances of both types whose confection can be very expensive, besides the annotation of these corpora with linguistic and semantic features, etc. for the training process (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 *instances of machine translationness*. We have coined the term machine translationness (Mtness, from now on) 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 their own output and they do not foresee, when faced to two or more possible translation solutions, the impact of choosing one of them, whether the recipient would take its output as intelligible and well expressed or, on the contrary, hardly intelligible and even nonsense. This capacity distinguishes human and machine translators, being the former constant evaluators of their output as they are producing it and always hypothesing the reaction of the recipient at their decisions. For example, let's see *mueran de siete* ('they die of seven') and *salida quiere* ('departures wants'), which are the Catalan-Spanish MT translations of *morin de set* ('they die of thirst') and *sortida vol* ('departure flight') respectively. These translations are instances of Mtness because their generation by a Spanish native speaker is very improbable and, besides, the system has been incapable of foreseeing the reaction of the recipient at the decision of translating *set* as *seven*, and *volen* as *fly*, decisions human translators would not make just because they consider them absurd translations and they know the recipients would consider them so.

In order to find instances of Mtness, we relate the probability of generation of a piece of MT output by a fluent speaker with the number of apparitions in a representative corpus of the target language. On the other hand, we relate the reaction of the recipients at a translation solution with the expectancy of the recipients of finding it in a fluent text. The expectancy is inferred by comparing the number of apparitions of each possible translation solution in the representative corpus.

So we have focused on instances of machine translationness 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 TC₁, TC₂,...TC_n possible translations, TC_i is an instance of machine translationness if the number of apparitions of TC_i in a representative corpus of the target language is null or its number is overwhelmed by the number of results of any of the other possible solutions.

For practical reasons, we take all the web pages published in the target language as the representative corpus. So the number of apparitions of a chunk can be inferred by the number of web pages containing it according to a search engine, provided that the target language is widely present in the world wide web. No results means that the apparition of an MT chunk in a fluent target language text is highly improbable so it may be an instance of machine translationness, whereas a chunk with more than, say, 1,000 results is not so considered. For example, the chunk *vuelan esconder* 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, machine translationness 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 SVM Tagger for Spanish developed by Giménez and Márquez (2004) for the evaluation prototype (see section 4).

Creation of MT output chunks. The POS-tagged MT output is splitted 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 and express a relation between two concepts. So far we consider 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 will be called as A1 and A2 from now on:

- 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. In case there is a word TW whose corresponding source word SW may have a different translation TW', substitute TW for TW'.

(4)

Catalan	Spanish C	SW	TW	TW'	Spanish C'
Sortida vol (Departure flight)	Salida quiere	vol	quiere	Vuelo	Salida vuelo

So far we have stated 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 these resources are needed: a source and target language wordlists, with the form, lemma and POS tag for each word, and a list of pairs <source word, target word>, where 'target word' is the translation 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 machine translationness. In a way similar to the selection of translation candidates in (Greffentette, 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. In case there are no results, the chunk is put in a list of candidates to be instances of machine translationness. 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 of an alternative chunk overwhelms the number of results of the MT chunk, the latter is considered an instance of machine translationness. The instances of machine translationness are stored in a list.

Results comparison. The number of instances of machine translationness of system A or its latest version is compared to the number of instances of system B or its previous version. The fewer is the number the better is the system or the version. The lists of candidates to be instances of machine translationness of 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 machine translationness.

4. Evaluation method prototype

In order to test the feasibility of the method, we tried to find instances of machine translationness in the MT Spanish translations for the 500 Catalan segments prepared by the UOC Linguistic Service (see Section 2). The translations were performed by the open-source system Internostrum¹ because the resources of this system can be obtained freely so the Catalan and Spanish wordlists and the list of pairs <source word, target word> could be generated automatically. We chose the Catalan-Spanish direction because these languages are very close and, consequently, the instances of machine-translationness would stand out more sharply. From the 396 errors detected manually (unknown words, typo errors, bad agreement, etc.) we focused on the following error-types.

- **Bad interpretation of the sense 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 taken the numeral interpretation in the wrong way.

- **Homonym 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. That’s why *sortida vol* is translated as *sortida quiere* in (4).

- **Illegitimate word-to-word translation (11,4%)**

This covers improper translation of acronyms (e.g: translation of *memòria RAM* as *la memoria RAMO* which means literally ‘bouquet memory’), translation 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’), non-dropped prepositions in verbal complements (e.g: *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 apocoptation (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 ‘ser’ and ‘estar’ (0,7%)**

Es (is) can be translated both as ‘es’ or ‘está’, i.e. permanent vs. episodic ‘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 cause 61,2% of the errors detected. The rest is spanned among errors which can be easily detected by any word and grammar corrector such as typo errors (19,2%) and unknown words (10%), bad agreement in gender or verbal person (4,3%), contraction and syntactic phonology errors such as *de el* instead of *del* (of the) or *y hizo* (he/she did) instead of *e hizo* (0,7%). Lastly, 4,3% of the errors are varied and unsystematic errors. In Table 1 we show some instances of machine translationness detected by our method. The right translation for most of these instances has been found by selecting the alternative that overwhelms the MT chunk with most results.

Error Typology	Source Chunk	MT Chunk	MT Chunk Results	Alternative Chunk	Alternative Chunk Results
Bad interpretation of the sense of a source word	Morin de set (die of thirst)	Mueran de siete	0	Mueran de sed	164
	Dia sagnant (bloody day)	Jornada sangrante	0	Jornada sangrienta	32,100
Homonym confusion	Sortida vol (departure flight)	Salida quiere	61	Salida vuelo	310
	Sortir a sopar (go out)	Salir a cena	7	Salir a cenar	19,200

¹ www.internostrum.com

	for dinner) endeutament net (net debt)	endeudamiento limpio	0	endeudamiento neto	1450
No apocoptation	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-to-word translation	Fer el préssec (to be taken for a ride)	Hacer el melocotón	0		
	Memòria RAM (RAM memory)	Memoria RAMO	6	Memoria RAM	1,320,000
Improper use of ser-estar	El disc és plè (the disk is full)	El disco es lleno	0	El disco está lleno	398
	És previst d'arriuar (it is expected to arrive)	Es previsto llegar	0	Está previsto llegar	200

Table 1. Instances of machine translationness detected via web searches.

5. Discussion

As we have seen in the previous section, our method if combined with a spelling and grammatical corrector may detect more than 90% of the translation errors of our evaluation test, and most of Mtness instances correspondingly. 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. Apart from this, the detection of Mtness instances provide information that can be useful for developing a semi-automatic postediting module and also to set a strategy to improve the output of the system. The pair ‘instance of machine translationness- alternative with most results’ could be presented to the posteditors of machine-translated documents and they would accept or not accept the alternative. The accepted alternatives would be propagated throughout the document and be stored in a repository in order to perform automatic correction of machine translated documents. So, the costs are widely overcome by the benefits of the results obtained and the possibility of reusing them. This is why we present the method as a ‘cheap’ method.

Besides, the results of the evaluation are significant because the method is consistent to the idea that human evaluators detect features that characterise machine translations and they penalise translations with a high probability to belong to the machine class rather than to the human class. Yet, we are aware this method is intended to perform fast, on-the-fly evaluations in order to get a significant ‘first impression’ of the quality of the output, which, on some occasions, is enough for the purpose of the evaluation, and on other occasions, it is the first stage of a sounder analysis of the output if the evaluation purpose requires it.

However, 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 grammar 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 its 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 no matter if the label is absolutely correct or not. So, if a word is tagged, say, as a noun when it is properly a verb, it does not make the difference for our purposes if the assumed nominal complements are taken as verbal complements instead; in other words, if a semantic relationship between them is detected. For example, no matter if *sortida vol* is tagged as a noun phrase followed by a verb, or it is just labeled as a noun phrase. The evaluator will trigger the same query.

Secondly, as regards web pages, the mere apparition of a certain chunk is not always significant to determine its machine translationness or its non-machine translationness. For example, the Spanish wrong translation of *pla d'estudis* (‘study plan’ in Catalan) as *plano de estudio* is found in the Internet because it coincides with the Portuguese term. Besides, we must 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 apparitions of the proper translation alternative (e.g. *disco llevar* 63; *disco duro* 8540000) or do not appear when the chunk coexist with another chunk in a larger query (e.g. *lo nueve* ('the nine') 369; *lo nueve gobierno* ('the nine government') 0). However, we would like to stress that although we have presented the Web as the hugest representative corpus, we do not mean that another kind of corpus cannot be representative of a linguistic use depending on the evaluation necessities. For example, if we evaluate a test suite of exams machine translated at a university, the corpus can be the bulk of exams of all the subjects taught at this university. If the corpus comes from published documentation which implies that the documents underwent a postediting process, the problems we have just mentioned would not arise.

On the other hand, the lack of results in a representative corpus is not always a direct indication of a machine translation error. For example, a perfectly Spanish grammatical chunk like *mataron a Rigobert Mallafré* ('they killed Mr. Rigobert Mallafré') has no results because Rigobert Mallafré is an individual not referred to on any web page. We are considering to substitute proper nouns an NP with a very frequent nominal head, even a proper noun, with the sense 'human' or 'institution', etc, according to an online lexical database like Wordnet, and count the new query. For instance, we could substitute *Rigobert Mallafré* for *un policía* ('a policeman') and we would get 552 results that tell us the chunk is not an instance of Mtness. In case we could get the selectional restriction of the verb argument from an online lexical resource ('human' in this case), the proper noun would be straightly substituted by an NP containing a human noun head.

6. Conclusion and future work

The evaluation method presented is still too preliminary but the first results we have obtained are encouraging enough to keep on working on its full development. Contrary to other MT-evaluation proposals that do not use human reference translations and are based on the ability of a classifier to distinguish machine translation qualities that are not characteristic of human, 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 very cheap. So the expectation of performing an evaluation with significant results with very low cost in time and money is reasonable.

Apart from the economic advantages, the data obtained by applying this method can be reused for other purposes. The list of instances of machine translationness provides information about the drawbacks of an MT system and they are very useful for developers to improve its performance (microevaluation). Besides, Finally, the method could be adapted to test the linguistic quality of the pages published on the Web. For instance, by detecting instances of machine translationness we can know whether a web page has been translated automatically and has not been post-edited.

We will carry out a fully evaluation of the method proposed in the language pair already studied and in other language pairs. Besides, we will try to suggest new actions that generate automatically alternatives for MT chunks with more syntactic rather than lexical content. In order to do this, we are thinking of applying the method by which we obtained idiomatic alternatives of expressing a content from the Web as we presented in Moré, 2004 for an English grammar checker based on Internet searches. Finally, we will exploit the MT evaluation method for post-edition tasks.

References

- S. Corston-Oliver, M. Gamon i C. Brockett (2001). A machine learning approach to the automatic evaluation of machine translation. *Proceedings of the Association for Computational Linguistics*. Toulouse, France. pp. 140-14.
- M. Gamon, M., A. Aue, and M. Smets (2005). Sentence-level MT Evaluation Without Reference Translations: Beyond Language Modeling. *Proceedings of the 10th Annual EAMT Conference*. Budapest.
- J. Giménez and Ll. Márquez (2004) SVMTool: A general POS tagger generator based on Support Vector Machines. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.
<http://www.lsi.upc.edu/~nlp/SVMTool/>

- G. Grefenstette, G. (1999). The www as a resource for example-based mt tasks. Machine Translation Task, Proc. Of Aslib Conference on Translating and the Computer. London
- A. Kulesza i 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.
- J. Moré, S. Climent, and A. Oliver (2004). A Grammar and Style Checker Based on Internet Searches. *Proceedings of the LREC2004*. Lisbon.
- K. Papineni, S. Roukos, T. Ward i 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.