

Gathering Empirical Data to Evaluate MT from English to Portuguese

Diana Santos

Linguatca, Oslo, SINTEF ICT
Pb 124 Blindern, 0314
Oslo, Norway
Diana.Santos@sintef.no

Belinda Maia

Fac. de Letras de Univ. do Porto
Via Panorâmica, s/n, 4150-564
Porto, Portugal
bmaia@mail.telepac.pt

Luís Sarmento

Linguatca, Porto, FLUP
Via Panorâmica, s/n, 4150-564
Porto, Portugal
las@letras.up.pt

Abstract

In this paper we report on an experiment to gather quality analyses from several people, with a view to identifying problems and reaching consensus over (machine) translation from English to Portuguese. We start the paper by showing how this project is part of a larger framework of evaluation campaigns for Portuguese, and suggest the need for amassing consensual (or at least compatible) opinions. We describe the various tools (Metra, Boomerang, and TrAva) developed and explain the experiment, its results, shortcomings and lessons learned. We then present CorTA, a corpus of evaluated translations (English original, and several automatic translations into Portuguese) and make some remarks on how to use it for translation evaluation.

Introduction

Let us begin by stating that the issue of evaluating translation is not new and is extremely complex (see e.g. Bar-Hillel, 1960). Machine translation (MT) evaluation has a long history, starting with the ALPAC (1966) report, which was extremely important for MT and NLP in general. However, we should also like to draw attention to two interesting facts: translation seems to remain one of the most popular NLP applications, and its output is judged by laymen in a way that no other complex intellectual activity is: while ordinary people would not think of criticizing a legal document written by a lawyer, an experiment designed by a physicist, or a diagnosis performed by a doctor, no one refrains from judging and criticizing the output of such a complex craft (or art) as translation.

In fact, translation is an interesting area because most people have strong opinions about the quality of particular (mis)translations (as opposed, for example to assessing the quality of IR results or abstracts). However, in most cases, it is remarkably difficult to elaborate objective criteria with which to classify, praise or reject specific translations. The work described in the present paper is an attempt to assess some of these analyses in a form that will later allow us to make generalizations.

Linguatca's efforts to start joint evaluation activities in the field of the processing of Portuguese, defined in the EPAV'2002 and Avalon'2003 workshops, selected three main areas¹: morphosyntax, leading to the first *Morfolimpiadas* for Portuguese (Santos *et al.*, 2003, Santos & Barreiro, 2004); information retrieval, with resource compilation (Aires *et al.*, 2003) and participation of Portuguese in CLEF (Santos & Rocha, forthcoming); and machine translation (MT), reported here and in Sarmento *et al.* (forthcoming).

It should be noted that these are radically different areas with different challenges and different interested participants. For MT, despite projects initiated in Portugal and in Brazil, the Portuguese/Brazilian developing

community has, on the whole, had very little impact on the outcome of current commercial systems, and specifically those available on the Web. However, and given that Portuguese is a major language in terms of the number of native speakers, there are plenty of international systems that feature translation into and from it, and there are many users of such systems worldwide. It was therefore thought that the best (initial) contribution that a Portuguese-speaking and Portuguese-processing community could offer was the identification of the specific problems (and challenges) posed by translation into Portuguese or from Portuguese. (We started with English as the other language.)

First, we thought about gathering test suites (of the translational kind of King & Falkedal, 1990), but in the initial process of discussing which phenomena should be extensively tested, there arose a more general concern with evaluating which kinds of problems were more obvious (and could also be consensually labeled) which led to the work described here.

The Porto node's concern with users in a language and translation teaching environment, and its close connection with the teaching activities at the Arts Faculty of the University of Porto, provided an excellent testbed for testing the possibility of collecting (machine) translation evaluations during the study programme. The pedagogical objective was to increase future translators' awareness of MT tools and encourage their careful assessment of current MT performance.

Gathering Judgements: TrAva

Our project had the double requirement of having trained translators with little formal knowledge of linguistics classifying the quality of the translation, and the need to create a classificatory framework that allowed comparison of examples, without making assumptions on the behaviour of specific MT systems. We have thus created a system for the empirical gathering of analyses called TrAva (*Traduz e Avalia*)², that has been publicized for general use by the community dealing with MT involving Portuguese, and whose continued use may supply new

¹ See <http://www.linguatca.pt/AvalConjunta/>, for information on the workshops and the three interest groups formed. ARTUR is the one for translation and other bilingual tasks.

² Available from <http://www.linguatca.pt/TrAva/>.

user requirements and functionalities, as well as larger amounts of classified data.

Due to the exploratory nature of this work, the use of the system by students and other researchers throughout the experiment has led to an almost continuous refinement of functionalities and several different versions. Although in the present paper we are restricted, for lack of space, to presenting only the current system, we must emphasize that the whole development proceeded bottom-up, and that the changes were motivated by the analysis of the input presented to the (previous versions of the) system.

TrAva is thus a system whose goal is to come to grips with some of the intuitively employed criteria of judging translation, by producing a relatively easy framework for cooperatively gathering hundreds of examples classified according to problems of (machine) translations.

From the analysis of the initial input to the system, it became clear that one should not rely on non-native competence to produce sentences to be translated, and thus we enforced the requirement that authentic English materials should be employed (and their origin documented, see Figure 1). Likewise, we required that only native speakers should classify translations, which means that so far we have only collected authentic English

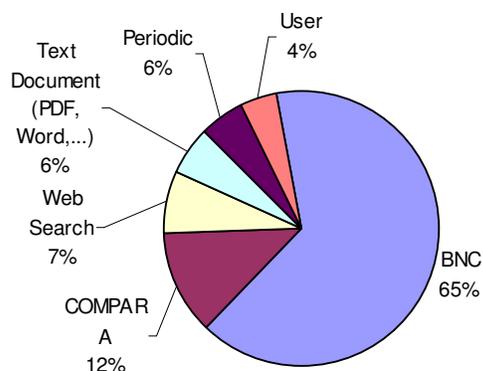


Figure 1 - Distribution of the origins of English sentences in TrAva (1270 sentences)

source language examples automatically translated into Portuguese and classified by Portuguese native speakers.

In order to be able to compare and gather large amounts of sentences with the same “classification”, and also to reduce subjectivity (or error) in the classification of the English text, we used the British National Corpus (BNC), Aston & Burnard (1996), and its PoS-tagging, as a first organization criterion. (Note that the students were also being taught to use the BNC in their translation education, so no additional training was required for the MT evaluation exercise.) The user is requested to indicate a sequence of PoS tags and classify the problems in the translation of this particular sequence, and not anywhere else in the sentence. One may submit the same sentence with a different target sequence, when additional interesting problems are observed, but ideally one should be considering each problem or structure in turn.

Due to the availability of Web-based MT engines, compared to systems that require acquisition, installation and/or format conversion, and given that we did not want to restrict the evaluation work to in-house members, but instead to offer it as a joint activity to the community

concerned with Portuguese language processing and even with MT, we chose to evaluate the performance of Web MT systems. As a preliminary step, two systems were developed: METRA (a meta-MT engine), <http://poloclup.linguateca.pt/ferramentas/metra/> and Boomerang, a system that sequentially invokes MT in the two directions until the same output is produced, <http://poloclup.linguateca.pt/ferramentas/boomerang/>.

They helped us identify problems and solutions to the engineering of invoking remote systems,³ and also gave us valuable insight into the relationships or dependencies among the seven MT engines involved. The final set used in TrAva contains the following four MT services: FreeTranslation, Systran, E-T Server and Amikai.⁴

A four-fold Classification Activity

The user of TrAva has, first, to decide which part of the sentence s/he is going to evaluate, and PoS-classify it. The text is then submitted to the four MT engines referred to above and the results are presented to the user, who reports on how many translations display problems in translating the selected part. Only then can the user engage in the most time-consuming (and complex) classification activity, namely to identify, using TrAva’s grid, the problems that appear in the translation(s).

Finally, and optionally, the user can also provide an alternative translation (this is encouraged), together with comments in free text. These comments have provided us with valuable input not only on several inadequacies of the current classification grids but also with feedback about the usability of the system. The alternative translation can also be considered a kind of classification (it may at least be used, in the future, as data for a re-classification, and for refining the grid).

A feature that may be difficult to understand is TrAva’s requirement that the user classify more than one translation at once, and thus it requires some explanation on our part: Our main wish is to identify cases which are difficult enough not to have been (totally) solved by any system yet, rather than compare the systems. One would expect to have problems that originate in the differences between English and Portuguese and that are not covered by current state of the art systems, such as questions, the translation of reflexives, modal verbs, homographs, complex noun phrases, etc, to mention just a subset of the problems investigated. So, we were expecting many translations to display the same or similar errors.

However, when it comes to a fine-grained classification of the problem, it appears that different systems often make different errors, and we are aware that it may be confusing for a user to try to classify all of them in one fell swoop.

Yet another Parallel Corpus: CorTA

One of the most relevant by-products of our experiment is CorTA (Corpus de Traduções automáticas Avaliadas), a corpus of annotated MT examples from English to

³ One has to deal with timeout, or “system not available”, with error messages, with excess length and consequent truncation, and – astonishingly – even sometimes with character codes and punctuation.

⁴ URLs are: <http://www.freetranslation.com>, <http://www.systransoft.com/>, <http://www.linguatec.de>, <http://standard.beta.amikai.com/amitext/>

Portuguese with non-trivial search possibilities. This novel resource has currently around one thousand input sentences (about 65% coming from the BNC) and, in addition to the usual search in parallel corpora like DISPARA (Santos, 2002), it allows for selection by kind of error and by translation engine. IMS-CWB (Christ et al., 1999) is the underlying corpus processing system.

CorTA is available at www.linguateca.pt/CorTA/, and is meant to grow at the rate required by the cooperative compilation of evaluations through TrAva. It is “frozen” in the sense that we do not plan to continue its development before October 2004, but until then we wish to receive feedback and gather more data, in order to assess what could be done and in which direction(s) it should be further developed.

This corpus is different in several ways from the one described in Popescu-Belis *et al.* (2002). Instead of a set of reference translations, it displays a set of (sometimes, correct, but usually incorrect) translations, which have not been hand-corrected, only hand-classified in relation to a subset of the problems they display.

Also, while the classification of Popescu-Belis *et al.*'s corpus is performed by a small group of experts (translation teachers), ours is cooperatively created by a set of people with little background, if any, in translation evaluation and is in principle open to any person who is a native speaker of one of the languages and knows the other well enough.

Although no numbers have been reported, we also expect the creation of such a corpus to be much more time-consuming than ours. On the other hand, their result will be a reference material, while ours, as it stands, can only be seen as a tool for empirical research in evaluation, translation, and human inter-agreement.

Lessons Learned

Although the system was initially created to allow cooperation among MT researchers, we soon learned that one cannot expect people to gather enough material for reliable research, without having some financial or other reward (such as project funding). Thus, if one wants people to consistently use a system whose primary goal is to provide data for later research, one has to employ students and/or people who may directly benefit from using it (such as those writing assignments).

So – as is, in fact, also the case in other kinds of empirical data gathering, such as software engineering (Arisholm et al., 2002) – one has to use students and not experts or translation professionals. In the case of TrAva, however, given that, as pointed out in the initial section, every one seems to have intuitions about translation quality, we believe that students of translation are expert enough when compared to “real” laymen.

Another relevant lesson is that very often a problem can be classified according to source-language, transfer/contrastive, or target-language criteria, and that this is a source of confusion to users of TrAva and consequently also of CorTA. For example, suppose the user was interested in the complex noun phrase *the running text mode* and one system had provided **o modo correndo do texto* (!). One could classify this erroneous translation as (English) attachment ambiguity wrongly analysed; (contrastive) incorrect resolution of ambiguous *ing*-form (adj-> verb); (Portuguese) wrong article

insertion/use, etc. All presuppose some model of how the system works – and may therefore be wrong – but by trying to guess the causes of the error, one may come to significant generalizations and, anyway, one cannot prevent people from thinking!⁵

So, while TrAva may seem flawed because different users may use different strategies to classify the problems, we believe it is also a strength that allows higher-level cause classification instead of simple objective correction. One is then able to look for all cases in the corpus that come from wrong PoS assignment regardless of the actual words or even the English patterns employed.

As the project developed, various other things became clear. For instance, we recognized the desirability of asking people to provide a good human translation, and the need to classify the MT output as acceptable in both Brazilian and European Portuguese.

Concluding Remarks

Obviously, the work we report here has never been thought of as an ultimate step in MT evaluation, but as a (maximally) unbiased pre-requisite for discovering a number of problems and for eventually producing a roadmap for MT into and from Portuguese.

We have not, at this stage, even tried to define metrics that could be employed to measure MT output, although we believe that CorTA could be a starting point for training automatic evaluators and for investigating the agreement with human intuitions about translation quality. There are a number of metrics and procedures used in MT evaluation (see Dabbadie *et al.*, 2002, for an overview), several of them making use of reference translations created by human translators, and specifying different translation goals (such as terminology coverage, NER handling, or syntactic correctness by counting the quantity of editing required). Because of their attempt at generality, they fail to consider the specific linguistic problems that the pairing of two particular languages poses. It is this language-dependent part that we want to address and to which we feel confident that the Portuguese-processing community can significantly contribute.

TrAva and CorTA are thus tools that allow everyone to look at specific problems of translation between the two languages and to suggest further ways to create representative samples (test suites) to test automatic translation per problem, instead of using “infamous” and irrelevant sentences from Shakespeare or the Bible or from the tester’s (lack of) imagination: “this is a test”, “hello, world”, and the like.

One should not forget that significantly and surprisingly good machine translation(s) can already be found as output of MT systems on the Web, and it is important to consider this in any reliable assessment. Although the judgments currently stored in TrAva are by no means representative, it is interesting to report that, in the process of testing probable sources of problems, users were partially happy in up to 66% of the cases, i.e., they considered 66% of the translations faultless regarding the phenomena under investigation, see Figure 2. CorTA can thus also be used as repository of solved problems (or of

⁵ On the contrary, instead of asking people to replicate machines, it would be more useful to ask them to think.

cases solved to a large extent), as well as of difficult cases to be used in future tests.

Finally, we must emphasize that, contrary to a test suite where the same lexical items are used many times in a controlled form, in CorTA, with TrAva, we can collect cases that display long-distance unforeseeable dependences and which would never even get addressed by more systematic means.⁶ Real running text is always preferable for evaluation of real systems in the real world, especially if one's intentions are not limited to evaluating a few already known phenomena.

Acknowledgments

The systems described in the present paper were implemented solely by the third author. The authors thank Fundação para a Ciência e Tecnologia for the grant POSI/PLP/43931/2001, co-financed by POSI. We thank all participants in the ARTUR interest group for their support and comments, as well as everyone who used TrAva and provided us with their feedback. We also thank the members of the Policarpo project at NILC and of TRADAUT-PT at FLUNL for making their own approaches to MT testing available to ARTUR. Finally, Anabela Barreiro helped shape and improve several

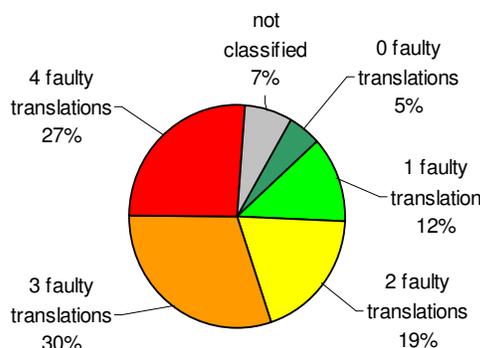


Figure 2 - Distribution of user classifications in TrAva (1270 sentences)

versions of TrAva with her feedback, and was responsible for devising several alternative input forms. The work reported here owes a lot to her being an enthusiastic power user of both TrAva and METRA.

References

Aires, R., Sandra Aluísio, P. Quaresma, Diana Santos & Mário Silva (2003). An initial proposal for cooperative evaluation on information retrieval in Portuguese. In Mamede et al. (eds.), *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003*, Proceedings, Springer, pp. 227-234.

ALPAC: Automatic Language Processing Advisory Committee (1966). *Language and machines: computers in translation and linguistics*. Division of behavioral

sciences, National Research Council, National Academy of Sciences, Washington.

Arisholm, Erik, Dag Sjøberg, Gunnar J. Carelius & Yngve Lindsjörn (2002). A Web-based Support Environment for Software Engineering Experiments. *Nordic Journal of Computing* 9 (4), 231-247, 2002.

Aston, Guy & Lou Burnard (1996). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bar-Hillel (1960). *Automatic Translation of Languages*. In D. Booth & R.E. Meager (eds.), *Advances in Computers*. New York: Academic Press.

Christ, O., Schulze, B. M., Hofmann, A., & Koenig, E. (1999). *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. University of Stuttgart, March 8, 1999 (CQP V2.2).

Dabbadie, M., A. Hartley, M. King, K.J. Miller, W.. M. El Hadi, A. Popescu-belis, F. Reeder & M. Vanni (2002). A Hands-On Study of the Reliability and Coherence of Evaluation Metrics. In M. King (ed.), *Machine Translation Evaluation – Human Evaluators Meet Automated Metrics, Workshop Proceedings, LREC'2002*, pp. 8--16.

King, M. & K. Falkedal. (1990). Using Test Suites in the Evaluation of Machine Translation Systems. In *Proc. of the 13th International Conference on Computational Linguistics (COLING)*, Helsinki, pp.211--216.

Popescu-Belis, A., M. King & H. Bentanar (2002). Towards a corpus of corrected human translations. In M. King (ed.), *Machine Translation Evaluation – Human Evaluators Meet Automated Metrics, Workshop Proceedings, LREC'2002*, pp. 17--21.

Santos, Diana (2002). DISPARA, a system for distributing parallel corpora on the Web. In Ranchhod, E. & N.J. Mamede (eds.), *Advances in Natural Language Processing (Third International Conference, PorTAL 2002)*, Springer, pp. 209--218.

Santos, Diana & Anabela Barreiro (2004). On the problems of creating a consensual golden standard of inflected forms in Portuguese. In *Proc. LREC'2004 (Lisbon, May 2004)*.

Santos, Diana, Luís Costa & Paulo Rocha (2003). Cooperatively evaluating Portuguese morphology. In Mamede et al. (eds.), *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003*, Proceedings, Springer, pp.259--66.

Santos, Diana & Paulo Rocha (forthcoming). CHAVE: topics and questions on the Portuguese participation in CLEF. *Proc. CLEF 2004 Working Notes (Bath, 16-17 September 2004)*.

Sarmiento, Luís, Belinda Maia & Anabela Barreiro (forthcoming). O processo de criação do TrAva e do CorTA. In Santos, Diana (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*.

⁶ A typical case is an NP erroneously torn between two different clauses in the source analysis, and therefore displaying lack of agreement in the translation. Independently of how the right error classification should be assigned in these cases, they clearly pinpoint errors that can only be observed and understood in a larger context.