Setting a Methodology for Machine Translation Evaluation

Widad Mustafa El Hadi – Université de Lille 3
Ismaïl Timimi - Université de Lille 3
UFR IDIST & CERSATES (CNRS UMR 8529)
Université Charles De Gaulle, Lille 3
BP 149, F-59 653 Villeneuve D'Ascq, France
mustafa,timimi@univ-lille3.fr

Marianne Dabbadie – LexiQuest
LexiQuest S.A.
Le Méliès
261, rue de Paris
F-93556 Montreuil Cedex
Marianne.Dabbadie@lexiquest.fr

Abstract

In this paper some of the problems encountered in designing an evaluation for an MT system will be examined. The source text, in French, provided by INRA (Institut National pour la Recherche Agronomique i.e. National Institute for Agronomic Research) deals with biotechnology and animal reproduction. It has been translated into English. The output of the system (i.e. the result of the assembling of several components), as opposed to its individual modules or specific components (i.e. analysis, generation, grammar, lexicon, core, etc.), will be evaluated. Moreover, the evaluation will concentrate on translation quality and its fidelity to the source text. The evaluation is not comparative, which means that we tested a specific MT system, not necessarily representative of other MT systems that can be found on the market.

Key-words

Black-box evaluation, Lexical Fidelity, Syntactic Fidelity, Non interactive MT evaluation, Terminology

1. Problem Overview

The object of this work is to set a methodology for non interactive machine translation evaluation on big corpora. We assume that the goal of the translation is a simple understanding of the original message (as it is for data mining for example). The goal of evaluation on a big corpus does not tend to exhaustive identification of incorrect translations as could be done manually on a small corpus. We did carry out some manual testing but with the objective of setting a rough methodology that may reveal in most cases non relevant translations on big corpora. This evaluation has been done manually on a small corpus but the methodology designed for this test is supposedly applicable to a larger corpus provided that the test is automated.

To carry out this work in rational conditions there was a need for:

(a) linguistic resources

(b) a set of procedures for screening the text through

(c) an MT System for output display

Given the issues we just pointed out, these viewpoints are not reductible. They are totally distinct focuses on the same object and must be analyzed in an autonomous way by referring to the theoretical sets of proposals, the techniques[1] and the practices on which they are based. The tools we used were a non interactive French / English MT System with a basic French/English dictionary that does not include specific terminology and two indexes: a French index and an English index of domain specific words for both languages, but these indexes were not aligned[2]. There was no post edition work on the target text or use of any translation memory. If we consider evaluation in this perspective we will have to respect these criteria. We will then first, explain and categorize the various choices within the framework of the previously-mentioned viewpoints.

A clear frontier must be set between verification and evaluation. Verification is a

[1] Within the framework of this evaluation we are not considering the theoretical groundings of the systems. It is a black-box evaluation.

[2] Aligning or pairing multi-lingual texts that are a translation of each other consists in making explicit the relations that exist between logical units of these texts. These units range from paragraphs and logical structure of a document, sentence, noun phrase, to words. The set of links (or pairs) is what is called an *alignment*. Both a multi-lingual corpus and an associated alignment are often call a *bitext*. Aligned texts, and therefore efficient alignment tools, for which there is an increasing need in many fields, such as lexicography or translation (Véronis *et ali*. 2000).

conformity check of system output to Software Requirement Specifications. According to the ISLE classification, the declarative evaluation on an MT system aims at measuring the ability of the system "to handle texts representative of an actual end-user". Moreover, it generally tests "for the functionality attributes of intelligibility (how fluent or understandable it appears to be) and fidelity (the accurateness and completeness of the information conveyed)".

These criteria (i.e. intelligibility and fidelity) precisely fall within the scope of the present work. Therefore we will measure syntactic and lexical fidelity of the target text. The two separate scores thus obtained will give the total score of the intelligibility of the translated text. We will deliberately leave semantic and pragmatic issues apart from this discussion, considering the automation of semantic representation has not yet yielded significant results to be used to evaluate the results of Natural Language Processing (NLP) systems such as machine translation, information retrieval or automatic text generation. It is important to note however that the use on an automatic semantic representations tool, provided that it were reliable enough to give an adequate representation of the input and the output of a system, would constitute a major advance for the evaluation of NLP systems.

2. Types of Analysis and Metrics

We have created a set of metrics to evaluate MT System syntactic and lexical correction rates and considering that this is also a manual study on a small corpus we decided to provide an exhaustive error analysis of non parallel data.

MT softwares can be classified according to whether they are based on resources of a linguistic or statistical nature. These systems normally share the following sets of features:

(i) Segmentation, a step which is usually considered as part of preprocessing operations on a text. It consists of two sets of operations:

(a) Dividing the text into separate sentences (paying special attention to the identification of typographical symbols and abbreviations, ..);

(b) Dividing the sentences into words (paying special attention to the processing of blanks, hyphens and so on);

(ii) morphological analysis (part-of-speech tagging);

(iii) syntactic analysis, taking into consideration word-category disambiguation, identification of noun-phrases and their functions;

(iv) unit extraction: category patterns; search and retrieval strategies for pattern extraction (domain specific terms and named entities);

(v) lexical analysis.

This does not mean that all softwares deal with these problems in the same way. For example, the morphological module can be constructed through a set of rules and/or a set of dictionaries; the syntactic module can be built either by parsing or by word category disambiguation, just to mention the two most opposite approaches ; the semantic module can be made more or less prominent. We are detailing the various types of analysis in the following sections adopting a black-box evaluation methodology as mentioned above.

2.1. Syntactic Analysis

We chose to count the number of *NPs* (noun phrases) and *VPs* (verb phrases) in source text and target texts, a first indication being given by non parallel data. *NP* is used in this paper to refer to both lexical NPs and non-lexical NPs. The former are distinctive entities requiring inclusion in the lexicon because their meanings are not unambiguously derivable from the meanings of the words that compose them (Justeson *et ali*. 1995), e.g. *fécondation in vitro* / in vitro fertilization; *Transfert d'embryon*/ Embryo Transfer; *banque de sperme*/sperm bank *; banque d'embryons*/ embryo bank, etc.

Lexical NPs are almost exclusively terminological. They are largely limited to those including adjectives and nouns only. They are often repeated in a text, a property which provides basis for their automatic identification, for instance (Justeson *et ali*. 1995). Whereas non-lexical NPs can include all types of parts-of-speech (determiners, adjectives, nouns, adverbs). The non-lexical NPs are known as GN when in contrast to GV as shown in the following examples taken from the text and analyzed by the syntactic tagger:

Sentence n° 1. *La production in vitro d'œufs fécondés et de jeunes embryons présente un intérêt majeur* (…). *La production* (..) and *un intérêt* are considered as non-lexical NPs, for instance.

Présente un intérêt majeur is considered as a VP in which an NP is embedded.

GN[Dét la] [N production Coord et] [GPrép[Prép de] [GAdj[Adj jeunes]][N embryons]]]]]i[GV présente [GN[Dét un] [N intérêt (…), translated in English by: "Production in vitro of fertilized eggs and young embryos presents a major interest (…)".

Moreover, lexical NPs (identified above also as GNs (*Groupe Nominal*)) when compared to non-lexical NPs: the former (lexical NPs) are subject to a much more restricted range and extent of modifier variations, on repeated references to the entities they designate, than are non-lexical NPs. This applies to variations in the omission of modifiers, in the insertion of modifiers, and in the selection among alternative modifiers. In contrast, omission of modifiers from a lexical NP normally involves reference to a different entity (Justeson *et ali*. 1995). Lexical NPs or "domain specific lexemes/terms" as we will call them in section 3.2. are far less sensitive than non-lexical NPs to other types of variations in the use of modifiers.

We remain clearly aware though, that a human translation from French to English might not necessarily generate the same syntactic analysis between source and target texts, given the gap generated by the translation of non parallel collocations or idiomatic expressions. But the translation made by a non interactive MT System that does not include any domain specific dictionary most of the time tends to provide a word to word translation. Therefore, on big corpora a sensitive difference in terms of quantity of NPs and VPs in source and target texts may then possibly reveal a wrong translation. A threshold could be fixed in an automated procedure including the use of a previously tested and reliable bilingual syntactic parser that would generate an output file providing NPs and VPs count. The use of finer grained criteria such as a count of adjectives or prepositional phrases could also be envisaged. Any overlap of this threshold might then be considered as an indication that MT system may have failed to analyze source syntactic structure and that therefore, these figures require further analysis. For the purposes of this study we used the LATL[3] bilingual syntactic parser[4] with a manual check and correction of errors. The metrics used to measure correction rate are detailed in the following subsection.

2.2. Syntactic Fidelity

To obtain a success rate we worked out the following rates:
1-(Number of target NP – source NPs) / Number of source NPs
And
1-(Number of target VP – source VPs) / Number of source VPs
Total Correction rate : (NP correction rate + VP correction rate) / 2.

2.3. Lexical Analysis

Checking lexical correctness includes the following subtasks:

- Polysemous words resolution: this is to check whether the system suggests the right target equivalent for a sense unit;
- Segmentation problems;
- Fluency problems (non idiomatic expressions – A detailed analysis is provided below in 3.2. but no numeric data will be given because we assume that MT goal in our study is limited to information).
- Domain specific terminology or lexical-noun phrases (NPs).

Let us assume that to one source meaning should correspond one target meaning (which is not linked to the number of words actually present in the text). A count of "meaning units" which can either be single words or collocations with several levels of granularity has been done on the corpus. The lexical evaluation has been done manually for the purposes of this study, notwithstanding the fact that an automated procedure should include a semantic representation tool on big corpora, a first indication being given by non parallel data, as for grammatical correctness. A sentence level fluency analysis will be carried out in this study but the reader should keep in mind, however, that lexical incorrectness at

[3] Laboratoire d'Analyse et de Traitement du Langage, University of Geneva.

[4] Syntactic analysis is one of the major components of a translation-oriented NLP which first applications began with MT. Analyses within the framework of an MT task can be seen as many sub-tasks which sum up the different relevant linguistic levels: morphological analysis, syntactic analysis (identifying noun and verb phrases and their functions) and finally, semantic analysis. Each of these sub-tasks can be in turn broken into smaller tasks: we can distinguish a) segmentation (identifying the word frontiers); b) lemmatization; c) tagging (identifying morpho-syntactic categories of each form), Abeille *et ali*. 2000.

sentence level cannot yet be worked out automatically. An indication of a possible lexical incorrectness on big corpora can be given though, by defining metrics that imply sense units count in source and target texts, calculated by a semantic tagger. As for the syntactic metrics a threshold should be fixed to evaluate semantic correctness on big corpora.

2.4. Lexical Fidelity

Let us assume that the intelligibility criterion includes the characteristics of the translation process, the output characteristics, the quality of the translation, and the quality of the target text as a whole. Our point of view is that the fidelity criterion tends to answer the following question : Is the text understandable ? Let us assume that to one source meaning should correspond one and only one target meaning (which is not linked to the number of words actually present in the text and has no impact on the string realization of that meaning, given the assumption that a semantic representation can give way to an unlimited number of reformulations but limits to one, though, the number of occurrences of a target meanings for a given source meaning). This allows us therefore to create a bijective relation between source and target sense units and to set a metric for fidelity that can be based on a count of the number of lexical units in the source text, as a referential figure. Success rate, precision and recall measures can then be worked out on target text.

After the syntactic tagging of source text, to obtain the number of sense units in source and target texts we applied the following metrics:

N° of words in text – N° of Determiners - N° of prepositions – N° of Coordination conjunctions.

To obtain a success rate:

(N° source sense units – total N° of wrongly translated sense units) / N° of source sense units

Total number of wrongly translated sense units = number of incorrect translations + unknown words + incorrect suggestions for polysemous word resolution.

We also calculated:

Lexical precision = number of relevant target sense unit / total number of target sense units

Lexical recall = number of relevant target sense units / total number of source sense units.

In order to work out the total quality of the output translation we set a final metric that gives in fact an average of correction rate and fidelity measures: intelligibility. The intelligibility metric can therefore be viewed as the quality of the translation as a whole. It may be worked out in the following way:

Intelligibility = average of correction rate + fidelity.

3. Manual Analysis of Output Errors

3.1.Syntactic Analysis:

A gap between source and target NPs was noted in 30 % of the cases. Further analysis of this phenomenon gave the following results. In most cases the gap is due to unknown words which involve a wrong part-of-speech categorization. This is explained by the fact that unknown words, whatever part-of-speech they may belong to are tagged as noun phrases. There are in fact 52 unknown words in target text, which is a great source of syntactic categorization errors and lowers the general quality of the output translation.

Errors can originate from a wrong part-of-speech categorization between source and target text: in sentence n° 10 for instance, we found the following: *Les conditions de capacitation in vitro différent selon les espèces (….).* différent is not identified as a flexion of the French verb *différer* (which means to be different from) but as the French adjective *différent* (different). As a matter of consequence, the output translation is a verbless sentence: "*The conditions of capacitation in vitro different according to sorts (…)".

A similar phenomenon appears in sentence n° 13 where a source preposition "entre" (which means between) is translated as a verb phrase in French: *La variabilité du taux de fécondation enregistrée entre différents éjaculats ou différents béliers (…)* is translated by : "*The variability of the rate of registered conception enters different éjaculats (…)".

Another source of errors is the wrong processing of coordination by MT systems, as can be shown in the following example (sentence n° 5): *La maturation cytoplasmique de l'ovocyte est nécessaire à la décondensation de la chromatine du gamète mâte et au bon déroulement des premières segmentations de l'oeuf.* The source sentence is translated by : "*Maturation cytoplasmique of the ovocyte is necessary for the décondensation of the chromatine of gamete masts and in the good progress of the first segmentations of the egg. The French collocation "*est* *nécessaire à* " (*et à ….et à*) s translated by "is necessary *for* (and in …) whereas the correct output should be "is necessary for …and for …".[5].

3.2. Lexical Analysis

Lexical analysis involves the following sub-sections: Granularity Levels: general language word level; polysemous word resolution; domain specific terminology and fluency problems.

These different levels of analysis can be illustrated by the following :

General language word level : this level of granularity corresponds to two categories (i) either simple lexical morphemes (lexemes), i.e. formed from only one element e.g. review or (ii) simple grammatical words such as *chez*, badly translated in English sometimes by *at or *to as shown in the following example: *chez les petits ruminants et les équins* is translated by**: "***at the small ruminants and the equines", sentence n°2.

Polysemous word resolution: for polysemous words the MT System we used often suggested various equivalents but some of them were not suitable :

For example: "*Milieu*" is translated by environment which is acceptable as a translation but the tool suggests another word *middle which is unsuitable in the context of sentence n° 7.

« *La co-culture du complexe ovocyte-cumulus avec des cellules de la granulosa permet d'améliorer l'aptitude au développement des oeufs FIV dans un milieu supplémenté en FSH (…),* »

"The co-culture of the complex ovocyte-cumulus with cells of the granulosa allows to improve capacity in the development of eggs FIV in an environment supplemented in FSH, (…)".

Fluency Problems: for the purpose of this article we mean by ***Fluency*** the capacity of the system to generate correct idiomatically formed expressions. We are limiting our examples to the good formation of domain specific terminology, mostly noun phrases. We noticed that a lot of translated English noun phrases contain prepositions (normally "of") however in English, empirically, only about 3% of terminological NPs contain prepositions[6] (most generally « of ») as shown in the examples hereafter[7]: "production in vitro" > *in vitro production*; "maturation of gametes" > *gametes maturation;* "transfer of embryos", > *embryo transfer*; "nuclear maturation in vitro"> *In vitro nuclear maturation*; "Maturation (cytoplasmique) of the ovocyte > *the ovocyte (cytoplasmique) Maturation*; delay of penetration of the ovocyte > *delay of the ovocyte penetration; "* The variability of the rate" > *the variability rate;* "The temperature of incubation" > *incubation temperature;* the rate of gestation is 50 % > *the gestation rate, etc.*

Domain specific terminology (lexical NPs)

The unknown simple word can be either a head or a modifier. We matched the list of the unknown expressions to the bilingual index considered as gold standard test material list[8]. We describe the results hereafter:

Simple unknown words and their status

a) The following words are simple terms (heads and modifiers) which are considered as domain specific terms (cf. INRA French Index): *capacitation, chromatine, cytoplasmique, micro-injection intra-cytoplasmique* (*cytoplasmique* is domain specific expression and part of a noun phrase acting as a modifier), *granulosa, polyspermie, transgenèse*, etc.

b) The following words are simple terms (heads and modifiers) but not necessarily domain specific (cf. INA French Index): *décondensation, éjaculats, épididymaire, équins,* inactivé, *ionophore, métaphase, oestradiol, oestrus* , *organelles, ovocyte, préovulatoires*, etc.

4. Results - Numeric Data:

4.1.1. Syntactic Metrics:

Source NPs	Target NPs	Source VPs	Target VPs
142	**184**	**38**	**40**

[5] The reader can refer to section 2.2. for the metrics we used to calculate syntactic fidelity.

[6] This is not the case for French NPs, but since we are evaluating the English translation we chose to limit our description to English.

[7] We suggest in italics the "expected" NPs translations

[8] This list is provided by the INRA. It can be associated to the test corpus. INRA (Institut National pour la Recherche Agronomique i.e. National Institute for Agronomic Research)

4.1.2. Correction Rate

NPs correction rate	VPs correction rate	MT System correction rate
0.70	0.95	0.83

4.2. Lexical Metrics:

Number of words in source text	Number of words in the target text	Total unknown words	Polysemous word resolution Suggested	Correct/ suitable suggestions	Number of incorrect translations	Number of source sense units	Number of target sense units
544	562	51	8 proposals	1	21	302	322

Fidelity : 0.73
Lexical recall : 0.83
Lexical precision : 0.76
Intelligibility : 0.78

The intelligibility figure, which gives an average of correction rate and fidelity measures reveals that the translation is understandable in 78 % of the cases. It is important to note that this measure corresponded approximately to the intuitive feeling left after reading the target text through. Viewing these results, however, together with the manual analysis of syntactic and lexical data leads to think that unknown words are generally a great source of semantic errors and wrong syntactic categorization. The MT System performance could probably then be improved by a thorough addition of specific terminology in the form of a custom dictionary pointing to the right meaning and part-of-speech for each domain specific word.

5. Perspectives & Further work

We will use the indexes provided by the INRA to create a specific dictionary in order to evaluate the impact of specific terminology when integrated to an MT System and after having run the system with a basic bilingual dictionary. These results will give us comparative data to evaluate the impact of the addition of a domain specific dictionary to an MT system and in particular, the influence of specific terminology over the total quality of the translated output.

6. References

Abeille, A. Blache, Ph. (2000). Grammaires et analyseurs syntaxiques. In : Pierrel, J.-M. éds. (2000). *Ingénierie des langues*, Traité 1C2 - Section informatique et systèmes d'information, p. 51-76.

EAGLES (1999). EAGLES Reports (Expert Advisory Group on Language Engineering Standards)http://www.issco.unige.ch/projects/eagles/ewg99.

ISLE (2001). MT Evaluation Classification, Expanded Classification. http://www.isi.edu/natural-language/mteval/2b-MT-classification.htm.

ISO (1999). Standard ISO/IEC 9126-1 Information Technology – Software Engineering – Quality characteristics and sub-characteristics. Software Quality Characteristics and Metrics - Part 1

ISO (1999). Standard ISO/IEC 9126-2 Information Technology – Software Engineering – Software products Quality : External Metrics - Part 2

ISSCO (2001) Machine Translation Evaluation : An Invitation to Get Your Hands Dirty!, ISSCO, University of Geneva, Workshop organized by M. King (ISSCO) & F. Reed, (Mitre Corporation), April 19-24 2001.

Justeson, J.S., Katz, S.M. (1995). "Technical Terminology: Some linguistic properties and an algorithm for identification in text", Natural Language Engineering 1(1), pp. 9-27.

Kilgarrif, A. (1998). "SENSVAL: An Exercise in Evaluating Word Sense Disambiguation Programs", in Proceedings. LREC, Granada, May 1998, pp. 581-588.

King (1999a) EAGLES Evaluation Working Group, report,http://www.issco.unige.ch/ projects/eagles.

King, M. (1999b). "ISO Standards as a Point of Departure for EAGLES Work in EELS Conference (European Evaluation of Language Systems), 12-13 April 1999.

Mustafa El Hadi, W. (1998). "Automatic Term Recognition & Extraction Tools: Examining the New Interfaces and their Effective Communication Role in LSP Discourse". In Mustafa El Hadi, Maniez, J. & Politt, S. éds Structures & Relations in Knowledge Organization, Proceedings of the Fifth International ISKO Conference, Lille, 25-29 Août 1998, pp. 204-212.

Sparck-Jones K., Gallier, J.R. (1996). Evaluating Natural Language Processing Systems: An Analysis and Review, Springer, Berlin.

Véronis, J., Langlais, Ph. (2000). ARCADE: évaluation de systèmes d'alignement de textes multilingues. In Chibout, K., Mariani, J., Masson, N., Neel, F. éds., (2000). Ressources et évaluation en ingénierie de la langue, Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF).