Quantitative Evaluation of Machine Translation Systems: Sentence Level

Palmira Marrafa[1] and **António Ribeiro**[2]

[1]Universidade de Lisboa
Faculdade de Letras
Group of Lexical and Grammatical Knowledge Computation (CLUL)
Avenida 5 de Outubro, 85 – 5º
P–1050–050 Lisboa, Portugal
Palmira.Marrafa@netcabo.pt

[2]Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática
Quinta da Torre
Monte da Caparica
P–2829–516 Caparica, Portugal
ambar@di.fct.unl.pt

Abstract

This paper reports the first results of an on-going research on evaluation of Machine Translation quality. The starting point for this work was the framework of ISLE (the International Standards for Language Engineering), which provides a classification for evaluation of Machine Translation. In order to make a quantitative evaluation of translation quality, we pursue a more consistent, fine-grained and comprehensive classification of possible translation errors and we propose metrics for sentence level errors, specifically lexical and syntactic errors.

Keywords

Machine Translation evaluation, translation quality metrics

Introduction

Much work has been done on evaluation of Machine Translation in the last ten years (see, for example, Balkan, 1991; Arnold *et al.*, 1993; Vasconcellos, 1994; White *et al.*, 1994; EAGLES, 1996; White and O'Connell, 1996; White, forthcoming). A common goal has been the design of evaluation techniques in order to reach a more objective evaluation of Machine Translation quality systems.

However, the evaluation of Machine Translation has been subjective to a great extent. ISLE (the International Standards for Language Engineering) aims at reducing subjectivity in this domain. It provides a classification of internal and external characteristics of Machine Translation systems to be evaluated in conformity with the ISO/IEC 9126 standard (ISO 1991), which concerns quality characteristics of software products. It assumes the need of a quantitative evaluation leading to definition of metrics.

However, that classification is not fine-grained enough to evaluate the quality of machine translated texts regarding the possible types of translation errors. Thus, in this work, we propose a more consistent, fine-grained and comprehensive classification at the individual sentence level. Our classification takes into account the internal structure of lexical units and syntactic constituents. Moreover, we propose metrics to make an objective quantitative evaluation. These metrics are based on the number of errors found and the total number of possible errors. The structural complexity of the possible errors is also considered in the metrics.

We selected some pertinent characteristics from the ISLE classification to measure the quality of sentence level translations, concerning lexical and syntactic errors, including collocations, fixed and semi-fixed expressions for lexical evaluation. As for syntactic errors, we built a typology of errors.

Our methodology was motivated by English, French and Portuguese parallel texts from the European Parliament sessions and also by translations obtained from two commercial Machine Translation systems.

In the next section, we present a motivation for the refinement of the taxonomy with some examples. After that, we summarise the classification and define the metrics used for the evaluation. In the following section, we discuss some previous work. Finally, we present the conclusions and the future work.

Motivation

ISO (the International Organisation for Standardisation) and IEC (the International Electrotechnical Commission) are the institutions which develop international standards. As for evaluation, an important standard is the ISO/IEC 9126 (ISO 1991). This standard distinguishes between internal characteristics which pertain to the internal workings and structure of the software and external characteristics which are the characteristics which can be observed when the system is in operation.

The ISLE Classification Framework for Evaluation of Machine Translation[1] provides a classification of the internal and the external characteristics of Machine Translation systems to be evaluated in conformity with the ISO/IEC 9126 standard.

Aiming to analyse Machine Translation systems from a user's point of view, we focussed on the external characteristics. We took the ISLE classification as a starting point for this evaluation.

Ideally an evaluation of a Machine Translation system quality should cover all the different parameters liable to be considered in a translation. However, this is a too complex task to be done in this early stage of our work. Thus, we decided to focus on the sentence level.

[1] http://issco-www.unige.ch/staff/andrei/islemteval2/mainclassification.html

The evaluation of this level deals with functionality, in particular accuracy, according to the ISLE classification:

2.2 System external characteristics 2.2.1 Functionality 2.2.1.2 Accuracy 2.2.1.2.2 Individual sentence level 2.2.1.2.2.1 Morphology 2.2.1.2.2.2 Syntax (sentence and phrase structure) 2.2.1.2.3 Types of errors 2.2.1.2.3.2 Punctuation errors 2.2.1.2.3.3 Lexical errors 2.2.1.2.3.4 Syntax errors 2.2.1.2.3.5 Stylistic errors

Fig. 1: Extract from the ISLE Framework

However, the characteristics listed above are not fine-grained enough for the evaluation. Moreover, the metrics proposed in the ISLE classification do not provide a sufficiently objective evaluation.

Scoring the Quality

We aim at quantifying evaluation as much as possible in order to reduce subjectivity. In this way, we have compiled a systematic list of lexical and syntactic properties which can be a source of translation errors at the sentence level. Refer to the Appendix for the main properties included.

This list is used to compute both the number of possible errors that can occur in a given sentence and the number of errors actually identified in that sentence. The translation quality score is computed with these numbers, as follows:

$$\text{Score} = 1 - \frac{\sum_{e=1}^{n} \#\text{identified error(e)} \times \text{weight(e)}}{\sum_{e=1}^{n} \#\text{possible error(e)} \times \text{weight(e)}} \qquad (1)$$

where *e* is the error type number. The score is weighted since we assume that not every error has the same impact on the translation quality. It seems fair to take into account how severe errors are. We claim that the weights of each syntactic dependency constraint should be determined in function of the probability of its occurrence. These probabilities are computed from an analysis of *corpora* as shown below:

$$\text{weight}(c) = \frac{\#\text{occurrences of constraint}(c)}{\sum_{C=1}^{n} \#\text{occurrences of constraint}(C)} \qquad (2)$$

The weight of constraint *c*, computed in this way, is equal to the weight assigned to error *e*.

We should stress that there are some syntactic phenomena which are more difficult to handle than others in some Machine Translation systems because of the expressiveness power of the systems' formalisms. However, in our approach, a syntactic phenomena which may be difficult to express in a Machine Translation system is not necessarily assigned a high weight just because it is more difficult. Its weight is based on their occurrence frequency in *corpora*. We are evaluating the translation quality and not the quality of Machine Translation system. We take it as a black box. That is, as mentioned above, we do not evaluate the system internal characteristics, according to the ISLE Framework.

Lexical errors are not as clearly definable. We believe that they should take into account how much they affect the understandability of a sentence. For example, 'fat ideas' seems to be more difficult to understand than 'big ideas'. We claim that WordNets can be used to weight the lexical adequacy. This weight may be computed by measuring the conceptual distance between the node which represents the expected lexical unit and the one which represents the translation obtained. In order to include this weight, we are currently working on a way to tune the formula of the metric presented in (1). To measure the conceptual distance we intend to extend the techniques described in Resnik (1999).[2]

Notice that determining the number of possible errors is not a trivial task since the identification of all constraints can be quite complex.

An example is given in Fig. 2 and Fig. 3 (both the original text and the Portuguese version of the text were extracted from a document of the European Parliament):

Original text	'Texts adopted by the Parliament'
Portuguese version	'Textos aprovados pelo Parlamento'
Translation by an MT System	'Textos adoptivo por Parlamento'

Fig. 2: Example of a Translation

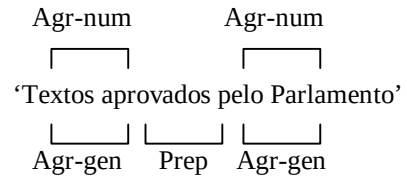

Fig. 3: Identification of Constraints.

Considering lexical and syntactic properties, these are the possible errors:

- Lexicon
  - Five tokens realised: 'textos', 'aprovados', 'por', 'o' and 'Parlamento';
  - One term: 'textos aprovados';
- Syntax
  - Agr-num: agreement–number between 'Textos' and 'aprovados';
  - Agr-gen: agreement–gender between 'Textos' and 'aprovados';
  - Prep: preposition selection: 'aprovados' selects 'por';
  - Order: four[3] wrongly ordered tokens;

[2] For an alternative approach, see Agirre and Rigau (1995).

[3] For *n* tokens, the highest number of token order errors is *n*–1, which happens when all tokens were reversed. We assume that

- Agr-num: agreement–number between ‘o’ and ‘Parlamento’;
- Agr-gen: agreement–gender between ‘o’ and ‘Parlamento’;

- Contractions: ‘por’ (‘by’) + ‘o’(‘the’) = ‘pelo’

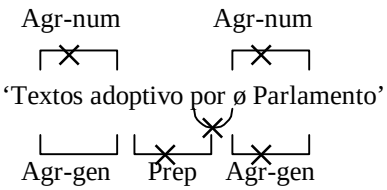

Fig. 4: Identified Errors

- Lexicon
  - One unrealised token: ‘o’;
  - One wrong token: ‘adoptivo’ instead of ‘adoptado’ (*co-occurrence restriction* violation);
  - One wrong term: ‘textos adoptivo’ (the European Institutions adopted the translation ‘textos aprovados’ for ‘texts adopted’);
- Syntax
  - Agr-num: agreement–number between ‘Textos’ (plural) and ‘adoptivo’ (singular);
  - Prep: preposition selection: ‘aprovados’ selects ‘por’;
  - Agr-num: no agreement–number between ‘o’ and ‘Parlamento’;
  - Agr-gen: agreement–gender between ‘o’ and ‘Parlamento’;
- Contractions: ‘por’ (‘by’) + ‘o’(‘the’) = ‘pelo’

The total number of possible errors found in this short example amounts to 16. This gives an idea of how hard the identification of possible errors in a text may be.

A Simpler Approach

Metrics strictly based on the total number of tokens and on the number of wrong tokens would obviously be much easier to compute.

Along these lines, Bangalore *et al.* (2000) discuss three metrics based on the number of insertions, deletions and substitutions needed in a generated string to obtain a reference string in the context of generation. Equation (3) shows the simplest one (Bangalore *et al.*, 2000, p. 3):

$$\text{Simple String Accuracy} = 1 - \frac{I + D + S}{R} \qquad (3)$$

where I is the number of insertions, D the number of deletions, S the number of substitutions and R the number of tokens in the string. This metric, which has already been used to measure quality of Machine Translation systems (Alshawi *et al.*, 1998), penalises twice words which are misplaced, as pointed out by Bangalore *et al.* (*ibidem*), because it counts this error as one deletion and one insertion. As a consequence, the number of insertions and deletions can be larger than the actual number of tokens. Should this be the case, the result of the metric may be negative. To avoid this, the authors treat the misplaced words separately in the formula by adding another variable (M) which counts the number of misplaced tokens.

$$\text{Generation String Accuracy} = 1 - \frac{M + I + D + S}{R} \qquad (4)$$

In spite of the improvement, this metric treats misplaced non-atomic constituents as several misplaced tokens. Thus, the authors recognise the need of including constituency criteria in the design of the metrics. As a matter of fact, creating discontinuities in constituents should be more penalised than scrambling constituents because the level of unacceptability is higher in the former case than in the latter. For example, ‘*Texts by adopted the Parliament*’ seems worse than ‘*by the Parliament texts adopted*’. Bearing this in mind, they suggest a third metric, called tree-based accuracy, which sums the score of the simple string accuracy metric, for atomic constituents and tokens, and the score of generation string accuracy metric, for non-atomic constituents. For this, each sentence is parsed to identify the constituents and its parse tree is compared to the tree of the reference string (the parsing is based on the Penn Treebank).

Nevertheless, this metric does not take into account the internal structure of constituents except for the linear order. As a consequence, whenever two errors occur in a token this approach just considers them as a single error. For example:

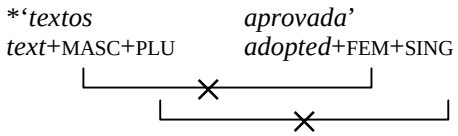

Fig. 5: Example of errors inside an NP (gender and number agreement)

In this example, we have two errors. However, the metric above just considers them as one, since it suffices the substitution of one token to correct it. This shows that we need to consider the internal structure of the constituents to identify and count all the errors in order to penalise them. Otherwise, some of them may not be penalised.

Our approach attempts to be more accurate, avoiding this problem. It considers the internal structure of the constituents, providing a more fine-grained typology of errors as presented in the Appendix.[4]

Conclusions

We believe that the approach presented in this paper is the right way to move towards a trustworthy evaluation of translation quality. Our proposal provides the means for an objective evaluation. It makes use of a fine-grained typology of errors which aims at dealing with boolean criteria. This highly reduces subjectivity.

one token resulting from two contracted or juxtaposed tokens counts as two distinct tokens. We do this because two non-contracted or juxtaposed tokens may be in the wrong order.

[4] Depending on the application, we can relax the granularity of the typology of errors. For example, specifier–noun agreement may not be relevant for gisting.

The typology of errors covers the main lexical and syntactic properties of sentences. The metrics weight syntactic errors taking into account the frequency of occurrence of the relevant constraints in *corpora*, and also lexical errors, adopting a measure based on the conceptual distance between nodes in a WordNet.

Future Work

At the moment, we are determining the probabilities of each type of syntactic error in order to establish the weights for the quantitative evaluation of this type of errors. In order to design objective evaluation metrics for the lexical adequacy, we are extending our evaluation metrics to include a measure of conceptual distance between lexical units.

In the next stage, we will extend this approach to text level evaluation, considering both text coherence and style.

Acknowledgements

We thank the anonymous reviewers for their useful comments and suggestions.

We also thank the organisers of the Machine Translation Evaluation workshop which took place in Geneva, Switzerland, in April, for providing the conditions which stimulated this work: Maghi King, Florence Reeder, Andrei Popescu-Belis and Sandra Manzi. Finally, we thank John White, Andrei Popescu-Belis and Sandra Manzi for providing us documentation.

References

Agirre, E. and Rigau, G. (1995). A Proposal for Word Sense Disambiguation Using Conceptual Distance. In Proceedings of the First International Conference on Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria.

Alshawi, H., Bangalore, S., Douglas, S. (1998). Automatic Acquisition of Hierarchical Transduction Models for Machine Translation. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Canada.

Arnold, D., Humphreys, R. and Sadler, L. (eds.) (1993). Special Issue on Evaluation of Machine Translation Systems. Machine Translation, volume 8, numbers 1–2.

Balkan, L. (1991). Quality Criteria for MT. In Proceedings of the Evaluator's Forum.

Bangalore, S., Rambow, O., Whittaker, S. (2000). Evaluation Metrics for Generation. In Proceedings of the International Conference on Natural Language Generation – INLG 2000. Mitzpe Ramon, Israel.

EAGLES (Expert Advisory Group on Language Engineering Standards) (1996). Evaluation of Natural Language Processing Systems: Final Report. Report for DG XIII of the European Commission.

ISO 1991. International Standard ISO/IEC 9126. Information technology – Software product evaluation – Quality characteristics and guidelines for their use. International Organisation for Standardisation, International Electrotechnical Commission, Geneva, Switzerland.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. In Journal of Artificial Intelligence Research, volume 11.

Vasconcellos, M. (ed.) (1994). Machine Translation Evaluation: Basis for Future Directions. Proceedings of a Workshop sponsored by the National Science Foundation, San Diego, California, USA.

White, J. (forthcoming). How to Evaluate Machine Translation Systems. In H. Somers (ed.), Computers and Translation: A Handbook for Translators, John Benjamins, Amsterdam, The Netherlands.

White, J., O'Connell, T. and O'Mara, F. (1994). The ARPA Machine Translation Evaluation Methodologies: Evolution, Lessons and Future Approaches. In Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia, Maryland, USA.

White, J. and O'Connell, T. (1996). Adaptation of the DARPA Machine Translation Evaluation Paradigm to End-to-End Systems. In Proceedings of the Second Conference of the American Association for Machine Translation – AMTA-96, Montreal, Canada.

Appendix

This is a summary of linguistic aspects to be taken into account for error identification.

1. Lexicon

Semantic Selection Restrictions
Collocations
Idiomatic Expressions
- Fixed Expressions
- Semi-fixed Expressions

Terminology

2. Syntax

Noun Phrase Level

Agreement (gender, number, case):
- Specifiers
  - Determiners
  - Quantifiers
- Modifiers
  - Adjectival phrases

Preposition selection restrictions
Order:
- Specifiers
  - Determiners
  - Quantifiers
- Modifiers
  - Adjectival Phrases
  - Prepositional Phrases
  - Relative Clauses

Prepositional Phrase Level

Pre/Postposition selection
Order of Pre/Postposition and Noun Phrase

Relative Clause Level

Relative Morpheme
Mode
Agreement:
- Relative Morpheme and Antecedent (number and gender)
- Antecedent and Verb (number and case)

Preposition Order:

- within Relative Phrase
- Preposition Pied-Piping vs. Preposition Stranding

Order of Relative Phrase

Verb Phrase Level

Order of Complements
Complements Case
Selection of Prepositions
Agreement:

- Negations (including Negative Concord)
- Time and Verb (tense and aspect)
- Complex Predicates: Noun Phrase – Adjective (gender, number)

Verbal Form (Simple or Complex):

- Mode
- Tense
- Aspect

Sentence Level

Agreement:

- Noun Phrase and Verb Phrase (gender, person, number, case)
- Secondary Predication

Complementisers

Ellipsis, Null Constituents, Contractions and Juxtapositions

Morphological Variances
Juxtaposition

Punctuation