The Impact of Reference Quality on Automatic MT Evaluation

Olivier Hamon[1,2] and Djamel Mostefa[1]
(1) Evaluation and Language Resources Distribution Agency (ELDA)
55-57 rue Brillat-Savarin, 75013 Paris, France
(2) LIPN (UMR 7030) – Université Paris 13 & CNRS
99 av. J.-B. Clément, 93430 Villetaneuse, France
{hamon,mostefa}@elda.org

Abstract

Language resource quality is crucial in NLP. Many of the resources used are derived from data created by human beings out of an NLP context, especially regarding MT and reference translations. Indeed, automatic evaluations need high-quality data that allow the comparison of both automatic and human translations. The validation of these resources is widely recommended before being used. This paper describes the impact of using different-quality references on evaluation. Surprisingly enough, similar scores are obtained in many cases regardless of the quality. Thus, the limitations of the automatic metrics used within MT are also discussed in this regard.

1 Introduction

Language resources (LRs) are essential components in research and development of NLP systems. However, the production of most LRs is done by human beings and is therefore subject to errors or imperfections. The creation of LRs requires a quality assurance procedure that helps control their quality and make sure that they comply with the specifications.

The importance of validation criteria is even higher when it comes to evaluation, as reference LRs are used to measure system performance and, thus, quality. An evaluation must be done in a suitable qualitative framework and data used should be as good-quality as possible. Bearing that in mind, validation standards have been defined (Van den Heuvel et al., 2003) and resources should follow the specifications as close as possible for that purpose.

The problem also applies to reference translations in MT. Most of the automatic metrics used compare human reference translations to the output obtained from MT systems. Generally, more than one reference is used to get multiple translation possibilities (Papineni et al., 2001), but the evaluation of sentences depends highly on the human reference(s) translation(s) used. However, only a few studies have gone deeper into a definition of quality and have tried to detail how to evaluate it (Van den Heuvel & Sanders, 2006). N-gram metrics give scores that strongly depend on the reference and, thus, we wonder how much scores computed with a poor reference translation diverge from the ones computed with a high quality reference translation. This paper focuses on two issues: 1) how to validate the quality of a human translation; 2) study of the impact of the quality of reference translations on MT evaluations. The final objective behind this work is to find out to what an extent a validation is useful within the evaluation protocol. The building of reference translations is very time and money consuming, but the cost of validation should not be underestimated either (Fersøe et al., 2006).

2 Context

In our experiments, we used the material from the TC-STAR [1] second evaluation campaign (Mostefa et al., 2006) and the third one (Mostefa et al., 2007). For both campaigns, three language directions were used: English-to-Spanish (EnEs), Spanish-to-English (EsEn) and Chinese-to-English (ZhEn). Data came from European Parliament Plenary Sessions (EPPS) for EnES and

[1] http://www.tc-star.org

EsEn, Spanish Parliament Sessions (Cortes) for EsEn, and Voice of America (VoA) for ZhEn. Three kinds of input were considered: automatic transcriptions from Automatic Speech Recognition (ASR) systems, manual transcriptions (Verbatim) and Final Text Editions (FTE) provided by the European Parliament. This represents 14 sets consisting of source documents, reference translations translated twice by two different agencies, and translations obtained from MT systems. Each set contains around 25,000 words. Therefore, we had an overall set of 28 reference translations on the evaluations, directions and inputs from both years. During the campaigns, MT systems have been evaluated with automatic metrics such as BLEU (Papineni et al., 2001).

3 Validation

3.1 Guidelines

The quality of reference translations is considered in two ways. First, translation guidelines are given to the translators. Then, translated files are sent to a validation agency in order to check their quality according to the defined criteria.

Guidelines were produced within the TC-STAR project. They were discussed internally but also with the Linguistic Data Consortium (LDC) who has had the experience of producing many reference translations.

Translation agencies are informed about the quality control and extra attention is paid to the recommendations given for translation quality: meaning and style should remain as close to the original source documents as possible; no additional annotations should be added to the translation; capitalization has to be carefully respected; the translation of neologisms and unknown words should take into account the speaker's intention; date format should also follow the established conventions, etc.

3.2 Criteria and Procedure

For each reference translation of the three language directions, the Speech Processing EXpertise centre (SPEX)[2] validated 600 words from contiguous segments randomly selected. Translations were checked by professional translators, who classified errors into categories. Points are given to references, according to the penalty scheme presented in Table 1.

Error category	**Penalty points**	
	Year 2	**Year 3**
Syntactical	4	3
Deviation from guidelines	3	-
Lexical	2	3
Poor usage	1	1
Capitalization	-	1
Punctuation / spelling (max.10)	0.5	0.5

Table 1. Translation errors penalties.

In order to be considered valid, a reference translation must have less than 40 penalty points. A non-valid reference translation is sent back to the translation agency/ies to be proofread and improved with the help of an errors report.

3.3 Typical Errors

Most errors are lexical ones, followed by poor usage of the language. Syntactic and spelling category errors are considerably fewer. In terms on input type, the number of lexical, spelling and syntactic errors is higher for FTE than Verbatim. On the other hand, the number of errors for usage and deviation from guidelines (including global translation quality) is higher for Verbatim.

Likewise, general errors are more frequent for English-to-Spanish than for Spanish-to-English. Chinese-to-English produces many more errors, in particular lexical ones. Syntactic errors could be wrong placement of adjective, wrong choice of person for pronouns, wrong use of verb tense or use of adjective as noun. Deviations from guidelines do not offer a wide variety: word/part of sentence omission, proper nouns mistranslation or translation quality/odd sentence problems. Thus, they have been regrouped under the others for the 3rd year evaluation. Lexical errors show the widest variety, probably due to the specificity of the vocabulary: mistranslation of acronyms, wrong word order, missing plural, literal translation, bad terminology or approximation, wrong preposition or translation inconsistencies. Other errors are wrong punctuation or spelling errors.

All these errors will lower the quality of the reference translations, which would imply a biased evaluation. The aim of the validation is then to reduce, as much as possible, the impact of mistranslation in order to improve *a priori* the assessment of the automatic translations.

4 Results

The following format is adopted for each set: "Year/Data-Input_Direction", e.g. "3/EPPS-FTE _EsEn" refers to the third-year set on Spanish-to-English using the FTE input on EPPS data.

[2] http://www.spex.nl

4.1 Results of the Validation

On the overall 14 sets, 10 sets have been translated again and revalidated at least once (for either one or the two reference translations). Table 2 gives the scores of validation for each of these sets together with the Word Error Rate with its respective validated reference. The left-hand side number gives the result for the first reference and the right-hand side one gives the result for the second reference. The different lines for each set give results for the different types of reference (from intermediate to final), thus showing the evolution of their validation.

Set	Validation score		WER / final reference	
	Ref 1	Ref 2	Ref 1	Ref 2
3/EPPS-FTE_EsEn	59.5 18 **18**	104.5 73.5 **38**	12.5 -	8.2 1.3
3/Cortes-FTE_EsEn	43.5 34 **34**	120.5 70.5 **35**	6.2 -	5.1 1.2
3/Cortes-Verb_EsEn	54 **26.5**	67 **22.5**	0.5	0.3
3/VoA-Verb_ZhEn	130 53.5 **27**	129 37 **37**	24.2 6.3	15.3 -
2/EPPS-FTE_EnEs	23 **23**	31 **23.5**	-	0.9
2/EPPS-FTE_EsEn	18.5 **18.5**	33 **17**	-	2.6
2/EPPS-Verb_EnEs	59.5 **11.5**	17 **17**	0.7	-
2/Cortes-FTE_EsEn	42 **6**	61.5 9	0.2	5.3
2/Cortes-Verb_EsEn	69 **18.5**	54 **0**	20.7	4.7
2/VoA-Verb_ZhEn	84 **39.5**	38 **38**	17.0	-

Table 2. Validation scores of reference translations and WER between intermediate (upper line) and final references (bottom line) for the first reference (Ref 1) and the second one (Ref 2).

The mean score for the reference translations before any correction takes place is around 71, while after correction this is around 23. Thus, final translations are not perfect but their quality is sufficient to serve as reference. However, a maximum of 130 is obtained from the translation for Chinese-to-English, which seems to be more difficult than the other directions. WER was also computed between the non-validated translations and their corresponding validated versions. As it can be observed, are not necessarily very high and many WER values are below 1%.

4.2 Intermediate vs. Final Reference

When comparing the differences between the validation scores and WER for each translation, no correlation is found. The correlation coefficient between the score differences and the WER is around 58%. For instance, a score difference of 36 between non-validated and validated references corresponds to a WER of 0.2, while another difference of 26.5 corresponds to a WER of 6.3. There is no direct correlation between the quality of the references and the WER scores. Thus, *a priori*, the quality of reference translations has no impact on the WER, which could be extended to the scoring of MT systems. Indeed, if WER does not reflect in a precise manner the quality increase of a human translation, how can it be useful/reliable for scoring MT systems?

Figure 1 presents the correlation between these score differences and WER. It shows that quality is not necessarily well correlated with WER. The explanation is twofold: firstly, the improvement of a human translation does not necessarily imply many changes; secondly, WER does not reflect the quality of a translation accurately, as it does not seem to focus on essential language issues.

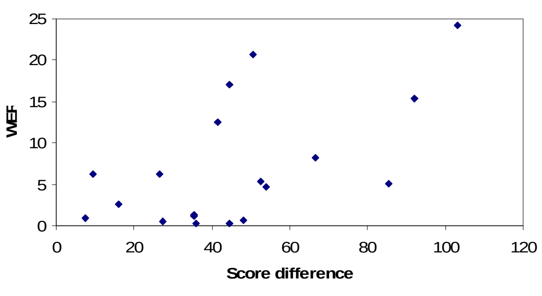

Figure 1. Correlation of score difference and WER between non-validated and validated references.

4.3 BLEU Results of MT Systems

BLEU scores have been computed for MT system output, using each set and each reference (whether validated or not). Then, either scores are quite identical or scores are slightly divergent. With the aim of studying this in detail, we assembled together the mean difference BLEU scores and the WER (against the final reference) for all the intermediate reference translations, as shown in Table 3.

The correlation coefficient between the absolute value of the mean difference score and mean of the WER is around 80%. Thus, the changes made into the references seem to have an impact on BLEU scores. However, given that quality is not correlated with WER, the absolute variation of the BLEU scores cannot be interpreted as a difference in MT system quality. It rather shows that comparing systems is the only plausible thing with BLEU, instead of evaluating systems in an absolute way.

Set	Mean diff. BLEU score	WER / final reference
3/EPPS-FTE_EsEn	1.18 -0.08	12.5 / 8.2 - / 1.3
3/Cortes-FTE_EsEn	0.67 0.035	6.2 / 5.1 - / 1.2
3/Cortes-Verb_EsEn	0.021	0.5 / 0.3
3/VoA-Verb_ZhEn	-1.71 0.001	24.2 / 15.3 6.3 / -
2/EPPS-FTE_EnEs	0.02	- / 0.9
2/EPPS-FTE_EsEn	0.24	- / 2.6
2/EPPS-Verb_EnEs	-0.05	0.7 / -
2/Cortes-FTE_EsEn	1.152	0.2 / 5.3
2/Cortes-Verb_EsEn	-2.21	20.7 / 4.7
2/VoA-Verb_ZhEn	0.08	17.0 / -

Table 3. Mean difference BLEU scores for each reference translation and WER between intermediate and final references.

4.4 Correlations of systems' evaluations

Correlations for BLEU scores were computed between 2 different-quality references. This allowed us to obtain 2 correlation coefficients for 2 non-validated references. For the correlation on scores, all coefficients are over 99%, so that even if scores increase or decrease, the distance between systems does not change. This is confirmed by the correlation on ranks, since the coefficients are between 96% and 100%. Thus, better reference translations could hardly distinguish MT systems in an easier way during evaluation.

5 Discussion and Conclusions

This work has used the BLEU metric to score MT system output and has demonstrated that the quality of reference translations does not have a clear impact on WER, also using n-grams. Even when using lower-quality translations, scores remain similar from one reference to another and important modifications of the human translation do not affect strongly the scores of the MT systems. This behaviour concerns all the languages tested, and remains the same regardless of the input or language used. However, we should not forget that the context of this experiment concerns actual automatic metrics. When reference translations have been modified, the impact on scores is not that clear, and even worse, this impact could be argued, to a certain extent, when the aim is to compare systems. Indeed, we also observed changes into scores when references were modified. Moreover, the quality of MT systems should not be ignored: if the overall quality of a system output is low, changes in reference translation will certainly have a lower impact on their scores.

Over the modification of the scores, the validation of the reference translation leads up to the validation criteria (although they are rigorously defined they are sometimes not very easy to apply by the validation team), the consistencies between agencies and translators (differences between reference translations show how the human translation quality may vary according to the translator) and some errors made by agencies (could be argued and validation can be difficult depending on the context, input, etc). Those points have to be carefully checked during a validation procedure and scores given by automatic metrics should be studied in agreement with the variation of the quality and validation.

Acknowledgements

This work was supported by the TC-STAR project (grant number IST-506738). We would like to thank SPEX, especially Henk Van den Heuvel and Eric Sanders, for the validation work. Our thanks also to Victoria Arranz for her help.

References

Fersøe H., Van den Heuvel H., Olsen S. 2006. *Validation of third party Spoken and Written Language Resources Methods for performing Quick Quality Checks.* Proceedings of Workshop “Quality assurance and quality measurement for language and speech resources”, LREC 2006, Genoa, Italy.

Mostefa D., Garcia M-N., Hamon O. and Moreau N. 2006. *Evaluation report, Technology and Corpora for Speech to Speech Translation (TC-STAR) project.* Deliverable D16.

Mostefa D., Hamon O., Moreau N. and Choukri K. 2007. *Evaluation report, Technology and Corpora for Speech to Speech Translation (TC-STAR) project.* Deliverable D30.

Papineni K, Roukos S, Ward T, and Wei-Jing Z. 2001. *BLEU: A Method for Automatic Evaluation of Machine Translation.* IBM Research Division, Thomas J. Watson Research Center.

Van den Heuvel H., Choukri K., Höge H., Maegaard B., Odijk J., Mapelli V. 2003. *Quality Control of Language Resources at ELRA.* Proceedings of Eurospeech, Geneva, Switzerland, pp. 1541-1544.