X-Score: Automatic Evaluation of Machine Translation Grammaticality

O. Hamon (1,2), M. Rajman (3)

(1) 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
(3) LIA - Ecole Polytechnique Fédérale de Lausanne Bât. INR - CH1015 Lausanne – Switzerland
E-mail: hamon@elda.org, martin.rajman@epfl.ch

Abstract

In this paper we report an experiment of an automated metric used to analyze the grammaticality of machine translation output. The approach (Rajman, Hartley, 2001) is based on the distribution of the linguistic information within a translated text, which is supposed similar between a learning corpus and the translation. This method is quite inexpensive, since it does not need any reference translation. First we describe the experimental method and the different tests we used. Then we show the promising results we obtained on the CESTA[1] data, and how they correlate well with human judgments.

1. Introduction

Most of the automatic methods for machine translation evaluation are used to compare system output with reference translation. Some of them, such as BLEU (Papineni & al., 2002) are based on comparison with many references, some others such as WNM (Babych & Hartley, 2004) attempt to use only one reference. All those automatic metrics attempt to avoid human evaluation which is quite expensive, but need nevertheless human generated reference translations. In fact those metrics are more semi-automatic measures than automatic ones.

The X-Score metric (Rajman and Hartley, 2001) is based on the distribution of elementary linguistic information within a text, such as morpho-syntactic categories, or syntactic relationships. The authors assume that this distribution of linguistic information is similar from one text to another within a given language.

With this automatic method, the X-Score is restricted to evaluate the grammaticality of a translated text; thereby we can rather speak about an evaluation of texts grammaticality applied to machine translation than an evaluation of translated texts.

Depending on the nature of the linguistic information selected to work with, the metric's precision will vary. For instance, working with syntactic dependencies will be much more precise that working with morpho-syntactic categories only.

Obviously, the primary advantage is to have no need to the source text, only the translated text is used, and there is no comparison between documents. As a black spot, we assume that the grammaticality of the source document is correct, in order to preserve at least the level of the grammaticality.

In this article, we propose an algorithm which computes a grammaticality score for a translated text. It is based on the following principle: first we compute a pattern of the linguistic information, such as morpho-syntactic categories or relationships within a fluency corpus representing the target language. Then we compute the number of occurrences of the morpho-syntactic categories within the learning corpus, in order to obtain a linear predictor to estimate the fluency score for any new input frequency list, computed from the frequency of morpho-syntactic categories within a text.

We present the results of an experiment which have been carried out on the data from the CESTA project (Surcin & al., 2005). Within the first evaluation campaign, five systems have been evaluated automatically and manually. This allows us to compare the results with automatic metrics such as BLEU, but also with scores from the human evaluation.

2. X-Score details

2.1. Overview

Only the translated text has been used for the evaluation. Indeed, we rely exclusively on the syntax of the target document, without taking into account either the semantic content or the syntax of the source document.

Before computing the X-Score, we first establish a typical representation of the grammaticality of the specified language. Then in a second step we compute the X-Score, where the language representation is applied to the translated text.

However, a pre-processing is required in order to build manually a fluency-scored corpus, which is composed of documents for which a fluency score is available. Fluency is used because it is held to be very similar to grammaticality (Rajman and Hartley, 2001).

2.2. Fluency Corpus

Before the learning stage, we need to do a pre-processing stage: build a corpus which contain document indexed by fluency. It is the only manually step during all the process of the metric.

The corpus contains several documents which a fluency score is given. The corpus used in the DARPA 94 evaluation (White & al., 1994) is similar to the one we used. In the DARPA 94 evaluation, human judgments were related to fluency, adequacy and informativeness scores. In our experiment, we assigned a fluency score for each sentence of the corpus, since fluency is held to be very similar to grammaticality.

The fluency corpus is only in the language of the target document, as grammaticality concerns only the translated document. The grammatical distribution has to be homogeneous in the corpus which is not necessarily very large. A thirty thousand words corpus seems indeed to be representative of a specific language.

[1] CESTA : Campagne d'Evaluation des Systèmes de Traduction Automatique, for Machine Translation Evaluation Campaign

For each segment of each document, a score of fluency is a mean of assessments assigned by a panel of several human judges. Then the mean of the fluency scores is computed for the whole document. Finally we obtain a corpus with fluency scores for each document. Those scores will allow establishing the grammatical model of the language regarding the linguistic information of documents.

2.3. Grammatical Model

The previously created corpus is used only during the learning stage, in order to represent the grammatical model. In this way for a defined language and a defined domain of application the model is established once and for all. Grammars being not frozen, there is not a perfect representation of it. It can notably depend on the applicative domain or the writing style.

For all documents of the corpus a morpho-syntactic tagger is applied to make out the corresponding morpho-syntactic tags and their occurrences.

Then we calculate the frequencies of each tag's type by document as:

$$f_k = d_k / N \quad (1)$$

where:

- f_k is the frequency of the tag k;
- d_k is the number of occurrence of k;
- N is the total number of tag's type.

Thereby, we obtain a vector of tags frequencies for each document, constituting a list of vectors.

As assumed by the X-Score's authors, the fluency score of a document is linearly dependant on tags' frequencies, which is written:

$$F_i = \sum a_k f_k(i)^2 \quad (2)$$

where:

- F_i is the fluency of the document i;
- F_k(i)is the frequency of the tag k in the document i;
- a_k the linear coefficient for the tag k corresponding to the document i.

The single unknown variant in this equation is the a_k linear coefficient, so we carry out a linear predictor in order to find this coefficient for all the documents such as the equation:

$$\sum(F_i - \sum a_k f_k(i)^2) \quad (3)$$

is minimal. Actually it implies that the coefficients are the same for all the documents of the corpus, such as the vector of the minimal coefficients is:

$$b = (X'X)^{(-1)}X'Y \quad (4)$$

where:

- X is the list of the frequencies' vectors;
- Y is the vector of the fluency scores;
- b is the vector of the minimal coefficients.

The b vector constitutes the list of the minimal coefficients and is used thereafter, for the evaluation of the translated document

To sum up the learning phase, the frequencies of the different categories of the selected linguistic information are computed and used to train a linear predictor able to compute a predicted fluency score for any new input frequency list. This linear predictor will then be used for evaluation.

2.4. Scoring

The second steps consist of computing a fluency score for the translated text using a linear predictor.

The vector of the minimal coefficients enables to make a projection of the fluency on the linguistic information of a corpus document, according to the distribution of the grammatical tags. This projection is a matrix of linear coefficients and is applied on the grammatical tags of the translated document.

As for a document from the corpus, a morpho-syntactic tagger is applied on the translated document then is computed the frequency for each tag with the equation (1). We obtain one more time a vector of tag frequencies.

The last step is to distribute the minimal coefficients from the learning stage on this vector in order to have the score of fluency for the evaluated document:

$$F_d = \sum a_k f_k(d)^2 \quad (5)$$

where:

- F_d is the fluency of the translated document d;
- F_k(d)is the frequency of the tag k in the translated document d;
- a_k the minimal linear coefficient for the tag k.

3. Experiment

3.1. The Fluency Corpus

Within the CESTA project a fluency corpus has been created. The original corpus contains five documents from the Written Questions and Answers of the Official Journal of the European Community (JOC), and four Arabic articles, of 270 segments.

In addition of the original references, we translated the English source documents into French by humans and automatic translators, and finally we obtained 221,686 French words and 2,778 assessments from 38 judges.

For fluency, the judges were asked to answer the question "is this text written in good French?" by giving a score on a 5-grades scale from "native French" to "non understandable".

For adequacy, they were asked to compare the meaning of the evaluated segment to that of a reference translation and score adequacy on a 5-grade scale from "whole meaning is present" to "nothing in common".

The judges were provided with guidelines before they started. These guidelines stipulate that they had to react as instinctively as possible and not spend more than 30 seconds on each segment.

To distribute the segments among the judges, all the submitted translations of the two tasks are merged as they all are in French. The segments are randomly divided out between the judges while assigning two judges for each segment. Finally each judge has around seventy segments to assess. The assessments have been done *via* a special web application specifically developed

for human judgments.

After the assessment the mean of the whole segments of each document have been calculated, therefore a fluency corpus was obtained, with a fluency score by document.

3.2. Evaluation data

The evaluation has been carried out by using a test corpus of 15 documents from the JOC corpus, and containing around 20.000 words segmented at sentence level. The documents were obviously different from those of the fluency corpus, but the grammatical distribution of the test corpus was logically close to that of the fluency corpus.

All documents were segmented at the sentence level, amounting to 790 English segments, and are not pertained to a specific thematic area, so their lexical coverage include a minimum of technical or restricted terminology. Four reference translations in French have been produced, one of them is the authoritative French version and the three other produced by translation agencies.

For the test these documents were randomly dispersed within "masking corpora" of more than 200,000 word which consisted of documents selected from the Economics and Diplomatic sections of the Financial Times newspaper.

For the human evaluation, each sentence has been assessed by two judges, as for the fluency corpus. Around a total of 140,000 words, 9,092 assessments have been done from 112 human judges who were recruited among students of French universities.

3.3. Parameters

Our metric remains experimental, so results are highly dependent on many parameters. In particular, it depends on the nature of the selected linguistic information, on the tool used to extract this information, and obviously on the fluency corpus.

In our experience we tried to investigate this metric for different types of linguistic information, and with different tools.To obtain the scores, we used the WinBrill tagger for French (http://www.atilf.fr), and we are currently re-computing the results with the Treetagger (Schmid, 1994) from the University of Stuttgart.

At this time we chose to evaluate the X-Score from a part of morpho-syntactic categories, with some variable composition of categories used. For instance we used the more relevant morpho-syntactic categories, such as noun, verb, etc. The punctuation tags have not been taken into account.

4. Results

4.1. Baseline Results

First it is important to notice that the better the X-Score is, the better the system is; but we attempt to privilege the overall ranking of systems rather than their scoring because of the difficulty to predict scores that correlate well with the Human judgment.

The Table 1 presents the baseline results for the Treetagger and Winbrill. We used all the available tags to obtain those results. It shows:

- The Human scores for Fluency;
- X-Score produced with the Treetagger;
- X-Score produced with Winbrill;
- The Human scores for Adequacy;
- Pearson's correlation coefficient for the both automatic measures correlated with Fluency and Adequacy scores of Human assessment.

At the top of the cells are show the scoring for each system; underneath are show the ranking for each system.

Systems	**Human Fluency**	**X-Score / Treetagger**	**X-Score / WinBrill**	**Human Adequacy**
System 1-EN	*0.459* **3**	*0.422* 3	*0.407* **3**	*0.5608* 3
System 2-EN	*0.419* **4**	*0.418* 4	*0.394* **4**	*0.5448* 4
System 3-EN	*0.353* **5**	*0.433* 2	*0.392* **5**	*0.4892* 5
System 4-EN	*0.511* **1**	*0.418* 5	*0.418* **2**	*0.6358* **1**
System 5-EN	*0.503* **2**	*0.435* 1	*0.420* **1**	*0.6080* 2
Corr. Flu.	-	*-0.3* -0.3	*0.93* 0.9	-
Corr. Ade.	-	*-0.25* -0.3	*0.95* 0.9	-

Table 1 - baseline results

The Treetagger's results are disappointing and have not been studied in more detail. Indeed, we choose to focus our study on the Winbrill tagger's results. Anyway we are re-computing the scores obtained with Treetagger in order to look at the possible problems. Even if the Treetagger seems not to work with our metric, the results obtained with the Winbrill tagger are very promising.

With the five systems presented above WinBrill tagger gives correlations of 0.93 with the human fluency score. For the system ranking, the X-Score strongly correlates with Human ranking, with 0.9.

With the two experiments below, we attempt to analyse in depth the X-Score with the Winbrill tagger.

As the Adequacy and Fluency scores are close, we strictly use the Fluency scores thereafter. In the same way correlation of we obtain for Fluency and Adequacy are close.

4.2. First experiment

For the first experiment, no normalization has been applied to the documents, what correspond to the results show above.

We detail the scores according to the tags and we choose four kinds of scores:

- A: all the tags
- B: only the relevant tags (adjective, adverb, noun, verb): ADJ ADV NN NNP SBC SBP VCJ VNCFF VPAR
- C : only the noun tags : SBC SBP NN NNP
- D : only the verb tags : ACJ APAR VCJ VPAR

ANCFF ANCNT ECJ ENCFF ENCNT EPAR VNCFF VNCNT ADJ1PAR ADJ2PAR VPAR

The Table 2 presents the results obtained with the non normalized documents.

Systems	A	B	C	D	Human Fluency
System 1-EN	*0.407* 3	*0.424* **1**	*0.443* 2	*0.396* 5	*0.459* 3
System 2-EN	*0.394* 4	*0.397* 5	*0.410* 5	*0.397* 4	*0.419* 4
System 3-EN	*0.392* 5	*0.411* 3	*0.448* **1**	*0.399* 3	*0.353* 5
System 4-EN	*0.418* 2	*0.421* 2	*0.429* 3	*0.402* 2	*0.511* **1**
System 5-EN	*0.420* **1**	*0.406* 4	*0.414* 4	*0.403* **1**	*0.503* 2
Corr. Flu.	***0.93*** **0.9**	*0.29* 0.3	*-046* -0.3	***0.61*** **0.5**	- -

Table 2 - Results on the non-normalized documents

Within the tests we realized that the input of the Winbrill tagger needs tokenized documents, therefore those good results are not so reliable. Irregardless, the B and the C evaluations are inconsistent regarding the Human results while they are expected to contain the most important information.

4.3. Second experiment

After the first experiment, we study the scores with the modified documents. We tokenized the entire documents, according to the specifications of the Winbrill tagger. For the automatic evaluation, we keep the same groups of tags. The Table 3 presents the results obtained with the non-normalized documents.

Systems	A	B	C	D	Human Fluency
System 1-EN	*0.400* 4	*0.419* 2	*0.446* **1**	*0.399* 3	*0.459* 3
System 2-EN	*0.388* 5	*0.391* 5	*0.410* 5	0.*396* 5	*0.419* 4
System 3-EN	*0.426* **1**	*0.420* **1**	*0.440* 2	*0.399* 4	*0.353* 5
System 4-EN	*0.409* 3	*0.413* 3	*0.430* 3	*0.401* 2	*0.511* **1**
System 5-EN	*0.419* 2	*0.402* 4	*0.416* 4	*0.402* **1**	*0.503* 2
Corr. Flu.	*-0.13* -0.1	*-0.18* -0.3	*0.43* -0.1	*0.65* **0.8**	- -

Table 3 - Results on the normalized documents

Obviously the new results are quite disappointing, as we were expecting better results than the previous evaluation.

The more noticeable ranking is that of the third system: with the A and the B evaluation the system is in first position, and in second position with the C evaluation. Only the last evaluation concerning the verbs improves the previous evaluation and close to the Human ranking with a ranking correlation of 0.8.

But except the D evaluation, all the correlations are very bad. We will attempt to find an explanation in the following sections.

4.4. Human vs Automatic

The last results presented here are the comparison between the X-Score and another MT evaluation metric: BLEU/NIST.

The Table 4 shows:

- cumulative 4-grams of BLEU scores and ranking produced using 4 reference translations and the true-case option;
- Winbrill scores and ranking produced using the tokenized documents;
- Winbrill scores and ranking produced using the non-tokenized documents;
- Human Fluency scores and ranking;
- Pearson's correlation coefficient for BLEU, Winbrill/tokenized and Winbrill/non-tokenized correlated with Fluency scores produced by the Human evaluation.

Systems	BLEU	Winbrill / non-token.	Winbrill / token.	Human Fluency
System 1-EN	*0.438* 4	*0.407* 3	*0.400* 4	*0.459* 3
System 2-EN	*0.465* 2	*0.394* 4	*0.388* 5	*0.419* 4
System 3-EN	*0.375* 5	*0.392* 5	*0.426* **1**	*0.353* 5
System 4-EN	*0.450* 3	*0.418* 2	*0.409* 3	*0.511* **1**
System 5-EN	*0.572* **1**	*0.420* **1**	*0.419* 2	*0.503* 2
Corr. Flu.	***0.68*** 0.5	***0.93*** **0.9**	*-0.13* -0.1	- -

Table 4 - comparison between the X-Score and BLEU

As we can see in the Table, BLEU scores and ranking are not as correlated with the Fluency scores and ranking as expected, but the Winbrill ranking is not correlated at all for the tokenized documents. On the other hand, the Winbrill ranking is strongly correlated for the non-tokenized documents.

To compare the automatic metrics, we also calculate the correlation coefficients:

- BLEU/NIST and WinBrill / non-tokenized are somewhat correlated: 0.70 for the both ranking and scoring;
- BLEU/NIST and WinBrill / tokenized are not correlated: -0.05 for the scoring, -0.3 for the ranking;
- Both Winbrill measures are not correlated either: 0.18 for the scoring while the ranking is nul.

4.5. Discussion

To conclude, even if the results are promising, they are rather uncertain.

As we can see in the Table 3, it is almost always the third system which obtains a very different ranking in relation to the Human judgments. If this system is discarded the correlation is widely better. The Table 5 shows the new correlation obtained without System-3:

Systems	**A**	**B**	**C**	**D**	**Human Fluency**
System 1-EN	*0.400* 3	*0.419* **1**	*0.446* **1**	*0.399* 3	*0.459* 3
System 2-EN	*0.388* 4	*0.391* 4	*0.410* 4	0.*396* 4	*0.419* 4
System 4-EN	*0.409* 2	*0.413* 2	*0.430* 2	*0.401* 2	*0.511* **1**
System 5-EN	*0.419* **1**	*0.402* 3	*0.416* 3	*0.402* **1**	*0.503* 2
Corr. Flu.	***0.92*** **0.8**	*0.49* 0.4	*0.23* 0.4	***0.93*** **0.8**	- -

Table 5 - correlations without System 3

The X-Score seems more competitive with this ranking. Therefore we have attempted to know more about the third system and its output. Let's take an example by considering the same French segment in output of two systems:

System-3 gives:

« Répond à donné par M Delors de la part de la Commission (10 le 1993 mars) »

which is not a really good translation. Furthermore the grammar quality is very poor (for instance "10 le 1993 mars" rather than "le 10 mars 1993", in French).

System-2 gives:

« Réponse donnée par M. Van Miert de la part de la Commission (le 28 janvier 1993) »

Here the sentence shows a higher grammar quality and is understandable.

The Fluency scores assessed by Human judges are 0.75 for System-2 and 0.125 for System-3. In that case, System-2 is obviously better than System-3. But the X-Score metric gives a score of 0.5359 for System-2, while it gives a score of 0.5570 for System-3. According to the X-Score, System-3 is better than System-2, but the human judges believe the opposite.

If we compare the both tagging for the segment, what is presented in the Table 6, it shows:

- words of the System-2 sentence
- tags corresponding to the words
- words of System-3 sentence
- tags corresponding to the words

The differences found in the comparison of the systems are presented in bold.

Sys-2 word	**Sys-2 tag**	**Sys-3 word**	**Sys-3 tag**
réponse	**SBC**	**répond**	**SBP**
		à	**PREP**
donnée	**ADJ2PAR**	**donné**	**SBC**
par	PREP	par	PREP
m	SBC	m	SBC
van	SBP	miert	SBP
		de	**PREP**
miert	SBP	camionnette	SBP
de	PREP	de	PREP
la	DTN	la	DTN
part	SBC	part	SBC
de	PREP	de	PREP
la	DTN	la	DTN
commission	SBP	commission	SBP
le	**DTN**	**28**	**CAR**
28	**CAR**	**le**	**DTN**
janvier	**SBC**	**1993**	**CAR**
1993	**CAR**	**janvier**	**SBC**

Table 6 - tags' comparison between the systems 2 and 3

At the top of the Table 6, we can see that System-3 contains more words than System-2. The corresponding tags are relevant tags, as names or adjectives (contrary to the preposition and determinant present in majority for the rest of the Table).

Furthermore, with the X-Score metric, the differences concern relevant tags versus non-relevant tags. For instance the third line compares "donnée" to "donné", however the corresponding tags are "ADJ2PAR" (adjective) and "SBC" (common name). The weights for this two grammatical tags are not the same.

Therefore System-2 is clearly privileged because it has more words which are more relevant regarding their tags.

The problem is not so easy to resolve, because it cannot be related to the metric.

A solution would be to compare the translated document with a reference translation and reduce the impact of the different number of words, introducing a weight. But at that moment the metric would not be entirely automatic, as is our purpose.

Another possibility is to establish this weighting from the source document on the assumption that the ratio between the tags of the source language and those of the target language are the same. There is again a limitation on this proposal, as languages have not necessarily the same grammatical structures. Anyway this could be predicted using two reference corpora: one in the source language and another in the target language.

In the fourth last line of the Table 6, the tags are exactly the same for both systems, but in a different order. System-2 is grammatically better on the four words though, but the automatic scores will be the same for the both systems.

An obviously solution is to improve the metric using bi-gram tags, as we plan to do in our further work.

5. Conclusion & Prospects

Even if the results are promising, we still need to improve our metric, as the results are very different depending on the parameters and the tools used. There are in particular great differences when using different taggers, and then the frequency lists are not the same.

We observed better results for systems which give more words in their output, even if it is a bad translation. There is also a problem with word order: two inverted words produce the same result, even if the syntactic structure is wrong.

Further works will be focused on the development of a new metric which will draw on the n-gram principle, at least with 2-grams.

6. Acknowledgements

This work is a part of the CESTA project, monitored by the University of Lille 3 (France) and ELDA.

We are very grateful to the help given by Tony Hartley.

At last we hold to thank the ATILF to have providing French data to use Winbrill tagger.

7. References

Babych B., Hartley A. (2004). Extending the BLEU MT Evaluation Method with Frequency Weightings. In *ACL 2004 Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, July 2004. pp. 622-629.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation, IBM Research Report RC22176 (W0109-022). In *Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 311-318.

Rajman M., Hartley A. (2001). Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. In *Proceedings of the Fourth Workshop on MT Evaluation, MT Summit VIII*, Santiago de Compostela, September 2001 (pp. 29--34).

Schmid H. (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, 1994.

Surcin S., Hamon O. , Hartley A., Rajman M., Popescu-Belis A., Mustafa El Hadi W., Timimi I., Dabbadie M., Choukri K. (2005). Evaluation of Machine Translation with Predictive Metrics beyond BLEU/NIST: CESTA Evaluation Campaign #1. In *Proceedings of the MT Summit X*, pp. 12-16 September, 2005, Phuket, Thailand. Pp 117-124.

White, J., T. O'Connell, F. O'Mara. (1994). The DARPA 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*, USA.

Zhang Y., Vogel S., Waibel A. (2004). Interpreting BLEU/NIST Scores: How Much Improvement? Do We Need to Have a Better System? In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, May 2004. pp. 2051-2054.