# Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores

**Martin Rajman[1], Tony Hartley[2]**

[1]EPFL Lausanne, CH
martin.rajman@epfl.ch
[2]ITRI, Brighton, UK
tony.hartley@itri.brighton.ac.uk

## Abstract

The main goal of the work presented in this paper is to find an inexpensive and automatable way of predicting rankings of MT systems compatible with human evaluations of these systems expressed in the form of Fluency, Adequacy or Informativeness scores. Our approach is to establish whether there is a correlation between rankings derived from such scores and the ones that can be built on the basis of automatically computable attributes of syntactic or semantic nature. We present promising results obtained on the DARPA94 MT evaluation corpus.

## Keywords

Fluency, adequacy, informativeness, automatic MT evaluation

## Introduction

The intrinsic quality of a translated text has two major attributes—its fidelity to the content of the source text and its naturalness, or fluency, as a text in the target language. Several studies have demonstrated a correlation between fluency and fidelity or attributes of fidelity, such as adequacy and informativeness (Carroll 1966, Nagao *et al.* 1985). This is a useful finding, since assessing fluency requires only monolingual evaluators whereas assessing fidelity requires bilingual evaluators—an even more costly process. Yet it remains the case that, for any one text, a large number of fluency judgments is needed in order to generalize away from their essentially subjective status.

It is therefore of central concern to explore whether there are any automatically computable scores that correlate well with the expensive, manually produced evaluations. Earlier work on the correlation with a system's ability to translate named entities (White *et al.* (2000) and our own work during the Geneva ISLE MT workshop 2001 on the correlation with precision and recall scores for Information Extraction tasks has shown the difficulty of this enterprise. Therefore, we attempted instead to predict, not the scores at the document level but rather the overall rankings yielded by these scores for the MT systems themselves.

## Manually computed scores

A-, F- and I-scores are respectively the Adequacy, Fluency and Informativeness scores produced during the DARPA94 MT evaluation exercise. From 1992 through 1994, DARPA conducted a series of MT evaluations as part of the Human Language Technology (HLT) initiative (White, et al, 1994). The largest of these included 100 newspaper articles in each of three language pairs (Spanish, French, and Japanese into English). Each pair was represented by several MT systems in various states of maturity, and also by two sets of human, professional translations. Each translation, in turn, was subjected to three separate evaluation types:

**Adequacy:** Subjects assessed the presence of correct meaning in the MT output—a *fidelity* measure.

**Informativeness:** Subjects answered multiple-choice content questions about each translated text, rather like a reading comprehension test—another *fidelity* measure.

**Fluency:** Subjects rated sentences in translated documents by the degree to which they were well-formed English, measuring *intelligibility.*

## Automatically computed scores

### C-score

The C-score is taken to measure the grammaticality of the translations. For any given document, the C-score is obtained as follows. First, a syntactic bracketing is produced for all the sentences in the document; this bracketing results from the analysis of the sentences by a stochastic context-free parser (the Slp-toolkit, developed at EPFL-LIA) trained on the Suzanne corpus (Sampson, 1994). To reduce the impact of Out-of-Vocabulary words, all surface forms that do not belong to the lexicon used by the parser are associated, for the analysis, with all possible open Parts-of-Speech. Since it often happens that a sentence does not get a full analysis (i.e. a bracketing spanning the whole sentence), all partial bracketings are retained. Then, for each sentence we compute, for every word subsequence W, the number of maximal bracketings covering W (i.e. the bracketings that cover the sentence but are not contained in any other bracketing covering another subsequence containing W). This enables us to compute the average bracketing coverage for the sentence (i.e. the average number of word contained in a maximal bracketing produced for the sentence). We assume that this number, which corresponds to the average size of the syntactic chunks that the parser was able to produce for the sentence, is intuitively representative of the grammaticality of the sentence (according to the parser). To enable the comparison of the average coverage between sentences of varying length, we normalize the raw values according to sentence length, to obtain values between 0 and 1. The C-score associated with a document is then simply the average of the normalized average bracketing coverages obtained for the sentences in the document.

## X-score

The X-score is taken to measure the grammaticality of the translations. For any given document, the X-score is obtained as follows. First, the document is analyzed by the Xerox shallow parser XELDA[1] in order to produce the syntactic dependencies for each constituent sentence. For example, for the sentence *The Ministry of Foreign Affairs echoed this view* the following syntactic dependencies are produced: SUBJ (Ministry, echoed); DOBJ (echoed, view); NN (Foreign, Affairs) and NNPREP (Ministry, of, Affairs).

On our corpus, XELDA produces 22 different syntactic dependencies, among which (the figure within brackets indicates the dependence occurrence frequency):

RELSUBJ[2501]: for example, RELSUBJ(hearing, lasted) in "*a hearing that lasted more than two hours*";

RELSUBJPASS[108]: for example, RELSUBJPASS( program, agreed) in "*a public program that has already been agreed on ...*";

PADJ[2358]: for example, PADJ(effects, possible) in *"to examine the effects as possible"*;

ADVADJ[433]: for example, ADVADJ(brightly, colored) in *"brightly colored doors"*.

After each document has been parsed, we compute its dependency profile (i.e. the number of occurrences of each of the 22 dependencies in the document). This profile is then used to derive the syntactic X-score using the following formula[2]:

**X-score = (#RELSUBJ+#RELSUBJPASS-#PADJ-#ADVADJ)**

## D-score

The D-score is held to measure how well the semantic content of a document has been preserved during translation. The underlying idea is to use a vectorial model for semantics (similar to those used in domains such as IR) and a large corpus of aligned translations. We measure whether the position of the source document in the semantic vector space defined by the part of the reference corpus in the source language is comparable to the position of the target document in the semantic vector space defined by the part of the reference corpus in the target language.

More precisely, for any document in the source language, we compute its semantic similarity with each of the reference document in the source part of the corpus. The similarity measure used in our experiments is the cosine similarity between the document lexical profiles (with the SMART ltn weighting scheme). Note that the matrix of similarities obtained is an indirect way of defining the position of the document in the vector space without requiring knowledge of its co-ordinates. We proceed in the same way for the translation of the document in the target language. This gives a matrix of similarities between the translation and the translations aligned with the original reference documents.

The hypothesis is then that the structure of the vector space built by the original source documents of the reference corpus is preserved by translation in the target language. Thus, this structure should be very similar to that of the semantic vector space built by the available translations of the reference documents in the target language. If we believe this hypothesis to be true (and we give some evidence for that below), then the following property is true:

If the semantic content of a document is well preserved during translation, then the similarity matrix associated with the source document in the source vector space should be very similar to the similarity matrix of the translation of the document in the target vector space.

The distance between the two matrices (i.e. the square root of the sum of the squared differences between the components) then intuitively serves as an indicator of the quality of the preservation of the semantic content after translation.

In order to have a measure (hereafter called the D-score) that varies in the same direction as quality (the higher the value, the higher the quality), we use a inverse function of the distance. In our experiments, we therefore used the following definition for the D-score:

$$\text{D-score}(D_{tgt}) = 1/(1+d(M_{src}(D_{src}),M_{tgt}(D_{tgt})))$$

$M_{src}(D_{src})$ (or $M_{tgt}(D_{tgt})$) is the similarity matrix for the source document $D_{src}$ (or the translation $D_{tgt}$) in the source (or target) semantic vector space.

As reference corpus, we used the JOC corpus containing 6729 documents comprising questions and answers to the European Community as published in the *Journal Officiel de la Communauté Européenne*.

### D-score: validation of the invariance hypothesis

To provide some evidence for this hypothesis (which is central to a correct interpretation of the D-scores), we carried out the following experiment. We split the reference corpus into two parts: a random sample of 500 documents used to test the hypothesis, and the remaining 6229 documents serving to build the semantic vector spaces. Then, for each source document D in the random sample, we computed the distances $d(M_{src}(D),M_{tgt}(D_i'))$ for all 500 translations $D_i'$ of the 500 source documents. If our invariance hypothesis is true, then, if $D_{i0}$' denotes the translation of the document D, we should have, for all $i \neq i_0$:

$$d(M_{src}(D),M_{tgt}(D_{i0}')) < d(M_{src}(D),M_{tgt}(D_i'))$$

Thus, the proportion of the documents in our sample for which the above property is true is indicative of the confidence we can have in our invariance hypothesis. We used a Student test to measure the confidence one can have in a high proportion of documents verifying the property, and found that, at a confidence level of 95%, more than 95% of the documents indeed verify it. This shows our hypothesis to be reliable.

## Correlations between the scores

The first part of our work was to analyze the existing correlations between the various scores (A-, F-, I-, C-, X- and D-scores). To do so, for each of the 6 systems for which the scores were available, we computed the Kandall rank correlation (with the Kendall tau coefficient) for each pair of scores.

Tables 1 and 2 summarize the values obtained. In each cell of the tables a result of the form ++(xx%) indicates that a (positive) correlation at a confidence level

[1] http://www.xrce.xerox.com/ats/xelda/

[2] Several formulae would have been possible for the X-score. We have selected one such that, if applied on the average dependency profile, it correctly predicts the average rank ranking.

(statistical significance) of at least xx% has been observed for all the systems. Empty cells indicate non significant correlations (i.e. correlations that can be rejected at a confidence level of at least 95%)

| Score | I | F |
|---|---|---|
| A | ++(50%)[3] | ++(60%) |
| F | ++(60%)[4] | |

Table 1: Rank correlations between A, F and I scores

In contrast to the results given for the A, F and I scores, for the automatically computed scores C, X and D, no significant correlation was observed for any of the pairs. All correlations could be rejected with a confidence level of at least 95% (except for the pair (C,X) for systems 1 and 4 -- with ++(60%) for 1 and ++(70%) for 4 -- and for the pair (C,D) for system 4 with ++(50%)).

| System | A:C | A:X | A:D |
|---|---|---|---|
| 1 | ++(20.0%) | ++(40.0%) | ++(95.0%) |
| 2 | | ++(60.0%) | ++(40.0%) |
| 3 | ++(70.0%) | | ++(95.0%) |
| 4 | ++(70.0%) | | |
| 5 | ++(90.0%) | | |
| 6 | | | |
| | F:C | F:X | F:D |
| 1 | | ++(30.0%) | ++(40.0%) |
| 2 | ++(90.0%) | | |
| 3 | | | ++(50.0%) |
| 4 | | | ++(10.0%) |
| 5 | ++(80.0%) | ++(70.0%) | ++(20.0%) |
| 6 | ++(70.0%) | ++(80.0%) | |
| | I:C | I:X | I:D |
| 1 | ++(60.0%) | ++(80.0%) | ++(20.0%) |
| 2 | ++(95.0%) | | |
| 3 | | | ++(10.0%) |
| 4 | | | |
| 5 | | ++(60.0%) | ++(10.0%) |
| 6 | | | ++(10.0%) |

Table 2: Rank correlations between the A, F, I and C, X, D scores

The correlations between the manually computed scores and the automatically computed scores are given in Table 2 and the names of the systems corresponding to the codes in table 3.

| Code | System Name |
|---|---|
| 1 | Human translation (reference) |
| 2 | Candide |
| 3 | Globalink |
| 4 | Metal System |
| 5 | Systran |
| 6 | XS |

Table 3: Identification of the MT systems

The main conclusion that one can derive from the obtained correlation figures is that there is no pair of scores for which a significant correlation can be observed for all the systems. As a consequence, it is not realistic to think of predicting the A, F, and I scores individually for each document as we were intending at the beginning of our experiments. However, it still possible to evaluate how well the overall ranking of the systems can be predicted on the basis of the automatically computed scores. Note that this is in fact what is expected from a MT evaluation: an overall ranking of the evaluated systems (the scores on the individual documents being less important).

The remainder of the paper therefore deals with the prediction of rankings on the basis of the automatically computed scores. The above correlation results are used as a guide for the choice of the predicting scores: to predict overall rankings based on F, we use the C score; to predict overall rankings based on A, we use the D score.

## Predicting overall rankings

The very first problem we face when trying to predict overall rankings is the production of the reference overall rankings that should be predicted. For the 6 systems evaluated in DARPA94, the raw evaluation material consists of the A, F, and I scores assigned by the experts to each translated document.

To decide how an overall ranking can be derived from the individual scores, we consider each set of scores assigned to the MT systems for their translation of a given document as one individual preference indication (or vote) over these systems. The DARPA94 corpus of 100*6 evaluated translations therefore represents a set of 100 hundred individual preference indications, and the overall ranking we are looking for is the one that optimally globally represents the set of individual preferences.

This is in fact a very hard mathematical problem well known to economists and political scientists (in the domain of voting theory for example) which has been shown (Arrow, 1963) to have no indisputable optimal solution.

However, the aggregation techniques that are often used include:

- ranking by average scores (average score ranking or ASR);
- ranking by average ranks (average rank ranking or ARR);
- ranking by average binary preferences (average preference ranking or APR).

We do not consider here other aggregation techniques, such as approval voting or multiple round voting schemes.

ASR has the advantage of great simplicity: for each of the systems, its scores are averaged over all the documents and the resulting average values are used to rank the systems. An important disadvantage of ASR is its limited robustness: the resulting ranking might be quite sensitive to outlying values

ARR, which by construction is much less sensitive to outliers, might therefore be a good alternative. For each

[3] with the exception of system 1, 2 for which no significative correlation was observed (++(30%) for 1 and --(5%) for 2).

[4] with the exception of system 1 for which no significative correlation was observed (++(30%)).

of the documents, the scores of the systems are first transformed into ranks and the average ranks obtained by the systems over all the documents are then used to produce the final ranking.

We mainly use APR as a second step to produce partial rankings when the total rankings produced as ASR or ARR appear too unstable.

## ASR and ARR rankings

For each of the 6 scores, we obtained the following ASRs and ARRs:

| | | |
|---|---|---|
| | A score | |
| ASR | 1 5 4 3 2 6 | ASR&ARR |
| ARR | 1 5 3 4 2 6 | 1 5(3 4)2 6 |
| | F score | |
| ASR | 1 5 2 4 3 6 | ASR&ARR |
| ARR | 1 5 2 4 3 6 | 1 5 2 4 3 6 |
| | I score | |
| ASR | 1 5 3 4 2 6 | ASR&ARR |
| ARR | 1 3 5 4 2 6 | 1(3 5)4 2 6 |
| | C score | |
| ASR | 2 5 3 4 1 6 | ASR&ARR |
| ARR | 2 5 3 4 1 6 | 2 5 3 4 1 6 |
| | X score | |
| ASR | 1 5 2 4 3 6 | ASR&ARR |
| ARR | 5 1 2 4 3 6 | (5 1)2 4 3 6 |
| | D score | |
| ASR | 5 3 4 1 2 6 | ASR&ARR |
| ARR | 5 3 4 1 2 6 | 5 3 4 1 2 6 |

where the ASR&ARR cells contain the (possibly partial) ranking combining ASR and ARR (the systems appearing within parentheses having no specific rank relatively to eachother)

Several observations can be made about the resulting rankings:

- ASR and ARR are identical (or very similar) in all cases; the few observed differences might be indicative of unreliable partial rankings (this point is studied in more detail below);
- system 6 clearly appears as the worst system in all evaluations;
- system 1 (the human reference) is indeed ranked first according to the manual scores but is ranked quite low by the mechanical scores (except maybe for the X score).

Another interesting point is to define a distance on the rankings in order to confirm or disconfirm the pairing predictions (i.e. which mechanical score should be used to predict which human score) we made earlier on the basis of the correlation results.

A possible distance on rankings is the Hamming distance which computes the number of pairwise differences. The distance definition can be extend to partial rankings by adding an average value of ½ for all the pairwise differences that involve a pair for which no preference decision has been taken.

The distances between the A, F, I and C, X, D rankings respectively, computed with these conventions, are given in tables 4 and 5. Table 4 contains the distances between rankings including the human translations. But as we have seen, the predictions are quite unreliable for the reference system (human translation) and it is therefore more sensible to produce the distance matrix derive from the rankings with element 1 removed. This distance matrix is given in Table 5.

| | C | X | D |
|---|---|---|---|
| A | 7.5 | 3.0 | 3.5 |
| F | 6.0 | 0.5 | 6.0 |
| I | 7.5 | 4.0 | 3.5 |

Table 4: Distance matrix including human translations

| | C | X | D |
|---|---|---|---|
| A | 3.5 | 2.5 | 0.5 |
| F | 3.5 | 0.0 | 0.5 |
| I | 3.5 | 3.5 | 0.5 |

Table 5: Distance matrix excluding human translations

Interestingly enough, the pairing predictions made on the basis of the correlation results are confirmed for the pairs (A,D) and (I,D), but disconfirmed for the score F for which the pairing (F,X) is preferred over the previously predicted pairing (F,C).

## Stability of ASR and ARR

As the average rankings analyzed in the previous section are derived from a limited number of documents (100), it is of great concern to have some evidence about the sensitivity of the produced rankings to the specific documents they have been derived from. One standard method for testing the stability of results derived form finite sets of data is bootstrapping. The general idea of the method is very simple: the original data set is used to produce a large number of random samples (called bootstrap replicates) of the same size N as the original data set. The random samples are used to produce the result for which we would like to estimate the stability, which will then be measured by a statistic computed on the set of bootstrap duplicates.

In our case, the random samples are simply built by N times randomly selecting among the original N documents. Notice that it often happens that in the bootstrap replicates, the same document is duplicated several times. To measure stability, we simply compute how many time the evaluated ranking is produced among all the rankings derived from the bootstrap replicates. In our experiments, for each of the scores, we produced 5000 bootstrap replicates of the original document set and computed the relative frequency in the resulting set of 5000 rankings of derived from the original document set. The results, given in Table 7, invite several observations:

- all the human rankings are substantially more stable than the mechanical ones;
- ARR rankings are indeed often more stable than ASR rankings;
- for all the rankings, except the ones derived (ARR) from the F and A scores, stability is a real concern as the produced ranking appears in fewer than 50% of cases among the replicates;

- because of its very low stability, the ranking derived from the C score seems to be unexploitable as such.

| | ASR | ARR |
|---|---|---|
| A | 0.5714 | 0.6490 |
| F | 0.4872 | 0.7760 |
| I | 0.4264 | 0.4128 |
| C | 0.0528 | 0.0850 |
| X | 0.3376 | 0.2138 |
| D | 0.3878 | 0.4706 |

Table 7: ASR and ARR ranking stability (5000 bootstrap replications)

As we have already observed when we compared the ASR and ARR rankings, an important part of the instability of the produced ranking comes from the fact that the data they derive from simultaneously substantiates not one single overall ranking but in fact several competing ones.

In order to analyze such phenomena, it is important to be able to produce the different rankings that are substantiated by a given scored document set. One possibility would be to explore the different rankings frequently produced during bootstrapping. This method however is not optimal as it does not allow us to make use of the fact that the several competing ranking probably share important common parts (i.e. subset of identical pairwise orders).

A better approach is to focus on the average (binary) preference rankings, as we see in the next section.

## Average preference rankings -- APR

Average preference rankings represent another way of producing an overall ranking as the synthesis of a set of several individual rankings. The method to produce an average preference ranking is quite simple. The individual rankings are first converted in set of binary comparisons on pairs. For each of the pairs i:j, we then compute how many times i has been rank higher than j and the resulting average ranking is the one corresponding to simple majority decisions for all the pairs. By convention, for each pair i:j, an associated value 1 (or -1) indicates that element i has a better (or worse) rank than element j. For partial rankings, a value 0 indicates that for the pair i:j, no ranking decision have been made.

The APR rankings are in fact far more complicated than they appear. Indeed, with the procedure described above, it is not guaranteed that the resulting set of binary decision effectively corresponds to a ranking.

Two types of problems may arise. First, some of the average binary decisions cannot be taken on the basis of a simple majority vote because the number of votes for each of the 2 possible decisions (1 and -1) are equal; in such a case, a partial ranking is produced and the corresponding decision value is set to 0 are already mentioned earlier.

Second, the resulting set of average binary decisions does not correspond to a ranking because of the fact that some transitivity relation is not verified (this is the well known Condorcet paradox stating that that the aggregation of rational --i.e. verifying transitivity-- preference sets can result in a irrational set of preferences (Saari, 1999)). For example the 3 individual rankings: [1 2 3 4], [2 3 1 4], and [3 1 2 4] result in an average set of decisions that cannot correspond to a ranking as it simultaneously requires that [1 2] and [2 3] (which implies, by transitivity, [1 3]) and [3 1].

One possibility for dealing with such situations is to relax the binary decisions that violate transitivity to unknown (value 0), again turning the set of binary decisions into a partial ranking. For example, if the above 1:2, 2:3 and 1:3 decisions are relaxed, we obtain the partial ranking {[1 4], [2 4], [3 4]} which can then be chosen as average preference ranking.

## Application to the MT scores

If we apply the method described above to the scores A, F, I and C, X, D, we obtain the following average (possibly relaxed) binary decision sets that are indicated in table 11 at the end of this document. They correspond to the following (possibly partial) rankings:

| | |
|---|---|
| A | 1 5 3 4 2 6 |
| F | 1 5 2 3 4 6 |
| I | 5(1 3 4)2 6 |
| C | 2 5(3 4)1 6 |
| X | 5 2 1 4 3 6 |
| D | 5 3 4(1 2)6 |

As we did previously for the ASR&ARR rankings, we can produce the distance matrix between the A, F, I and C, X, D scores (with and without System 1—see Tables 8 and 9).

| | C | X | D |
|---|---|---|---|
| A | 7.5 | 5.0 | 3.5 |
| F | 5.5 | 2.0 | 6.5 |
| I | 5.5 | 4.5 | 2.0 |

Table 8: Distance matrix including human translations

| | C | X | D |
|---|---|---|---|
| A | 3.5 | 3.0 | 0.0 |
| F | 1.5 | 0.0 | 3.0 |
| I | 3.5 | 2.5 | 0.5 |

Table 9: Distance matrix excluding human translations

Thus, if we sum up the pairing predictions produced by the different methods seen so far, we obtain Table 10.

| | C | X | D |
|---|---|---|---|
| A | 0 | 0 | 3 |
| F | 1 | 2 | 0 |
| I | 1 | 0 | 2 |

Table 10: Pairing predictions—summary

The A and I scores should therefore be predicted by the D score, and the F scores by the X score.

## Reliability of the APR rankings

As it was the case for the ASR and ARR rankings, the issue of reliability is an important concern for the APR rankings. However, the approach for measuring reliability is quite different. As the APR is build by deriving the average binary decision from the counts of the individual binary decisions, a statistical test can quite easily be used instead of the simple majority rule. More precisely, this corresponds to replacing the rule: "select a decision if the proportion of individual decisions it corresponds to is greater than a half" by the statistical test "select a decision if the proportion of individual decisions it corresponds to is significantly greater than a half". As we are dealing with proportions, we used a Student test. The level of confidence that can be associated with a produced APR is then the lower of the levels of confidence that were used to select the average binary decisions. Applying this method, we obtained the following results for the different scores:

| | | |
|---|---|---|
| A (89%) | [1 5(3 4)2 6] | [5(3 4)2 6] |
| F (83%) | [1 5 2(3 4)6] | [5 2(3 4)6] |
| I (89%) | [5(1 3 4)2 6] | [5(3 4)2 6] |
| C (83%) | [2(1 4) 6],[(3 5)6] | [2 4 6],[(3 5)6] |
| X (83%) | [5 2 4 3 6],[1 4] | [5 2 4 3 6] |
| D (97%) | [(3 5)4(1 2)6] | [(3 5)4 2 6] |

where the second (resp. third) column contains the ranking including (resp. excluding) human translations (system 1) and the percentages indicate the levels of confidence corresponding to the (possibly partial) rankings.

## Conclusions

We set out to find a way of scoring translated texts automatically—and therefore relatively cheaply—such that the scores assigned would correlate with human evaluations of these texts in terms of their Fluency, Adequacy or Informativeness—their F-, A- and I-scores. The automatically computed scores are intended to reflect the grammaticality of the translations (the C- and X-scores) and the degree to which they preserve the semantic content of the original (the D-score). We discovered, in fact, that there is no pair of manually and automatically calculated scores for which a significant correlation can be observed for all five MT systems considered. However, we did succeed in establishing a correlation between the (partial) rankings of the MT systems given by the F-, A- and I-scores and those given by the C- and X- and D-scores, such that the latter can reliably predict human rankings of MT system performance.

When we rank the five systems according to the A-, F- and I scores, Systran (5) is the best MT system on all scores; XS (6) is the worst system on all scores; Globalink (3) and Metal (4) are indistinguishable on all scores; Candide (2) performs better than Globalink and Metal on those attributes related to content (Adequacy and Informativness) but performs worse than Globalink and Metal for those attributes related to syntax (Fluency).

When we rank the five systems according to the the C-, X- and D-scores, we observe that the X score is the best predictor for the F score; the distance between the APR ranking produced on the MT systems by the X score and the "true" APR ranking derived from the F score is 0.5 (corresponding to a similarity of 95%).

Other measures of the quality of the predicted ranking are its precision and recall (i.e. the proportion of binary comparisons correctly predicted among all the binary relations predicted and the proportion of binary comparisons correctly predicted among all the binary relations in the true ranking). We have: Precision(X) = 100%; Recall(X) = 100%.

We recognize that this represents an upper bound, since the calculation of the X-score was tuned to the particular corpus. Further work is required on other, larger corpora to establish whether the current definition of the X-score is more generally applicable.

The D score is the best predictor for both the A score and the I score (which produce the same APR ranking); the distance between the APR ranking produced on the MT systems by the D score and the "true" APR ranking derived from the A score or the I score is 1.0 (corresponding to a similarity of 90%). In addition, Precision(D) = 100.0%; Recall(D) = 88.9%.

The D-score appears to be both reliable and robust, and we believe that it will yield similar results on other MT output data.

A final remark concerns the large discrepancy between the manual and automatic rankings of the human translations, which raises interesting questions about the sensitivity of human judges to documents written by humans. A possible interpretation of the fact that the human translations are ranked low by the mechanical scores could be that judges actually recognize human translations and rank them using more complex criteria, while for the MT output they apply simpler evaluation criteria that are then easier to predict on the basis of simple observable properties.

### Acknowledgements

We would like to thank Florence Reeder and John White for providing us with the DARPA94 MT evaluation data, as well as Antoine Rozenknop and Romaric Besançon for their very valuable help with all the experiments.

## References

Arrow, K.J. (1963). Social Choice and Individual Values, Wiley, New York.

Carroll, J.B. (1966). An experiment in evaluating the quality of translations. In: J. Pierce. Language and machines. Report by ALPAC. NASNRC, pp. 67-75.

Nagao, M., Tsujii, J. & Nakamura, J. (1985). The Japanese government project for machine translation. Computational Linguistics 11(2-3), pp. 91-109.

Saari, D. (1999). Explaining all three-alternative voting outcomes, Journal of Economic Theory. 87 (pp 313-355).

Sampson, G. (1994). The Susanne corpus, release 3. In School of Cognitive & Computing Sciences, Brighton (UK). University of Sussex, Falmer.