ACCEPT

SEVENTH FRAMEWORK PROGRAMME

THEME ICT-2011.4.2(a)

Language Technologies

ACCEPT
Automated Community Content Editing PorTal

www.accept-project.eu

Starting date of the project: 1 January 2012

Overall duration of the project: 36 months

Analysis of existing metrics and proposal for a task-oriented metric

Workpackage n° 9	Name: MT Evaluation
Deliverable n° 9.1	Name: Analysis of existing metrics and proposal for a task-oriented metric
Due date: 31 December 2012	Submission date: 21 December 2012

Dissemination level: PU

Organisation name of lead contractor for this deliverable: UNIGE

The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement n° 288769.*

Contents

Analysis of existing metrics and proposal for a task-oriented metric

Objectives of the Deliverable

In the ACCEPT project, automatic evaluation metrics will be used in the context of WP9, a core module covering the measurement of MT from both automatic and human perspectives.

The specific objectives of using automatic metrics in this WP are the following:

- Evaluate the impact of pre-editing rules on translation quality (**Task 9.1**)
- Evaluate the impact on post-editing rules on translation quality (**Task 9.2**)
- Compare post-edited RBMT with post-edited SMT (**Task 9.4**)
- Assess usefulness, adequacy and reliability of automatic metrics by determining correlation with human judgments (**Task 9.4**).

This report, whose aim is to summarize automatic metrics and outline requirements for a new task-oriented one, should be viewed as preliminary; the first task in WP9 (Task 9.1) began in Month 12, and detailed specification of a new metric will only be possible when pre-edited and post-edited data are available to give a clear picture of the problems that need to be addressed.

Specifications and requirements for evaluations relative to Task 9.1 (Months 12-18) are described in Deliverable D2.1 and will not be repeated here. Those for Tasks 9.2 (Months 12-24) and related Task 9.3 (month 18-24) will be defined in Deliverable D2.3 (Month 20) when post-editing rules are operational (as requested in the Description of Work). Tasks 9.4 (Months 12-30) will be planned later in the project in Deliverable D9.2.1 (Month 18).

Description of the metrics used

The field of machine translation has in recent years developed a number of automated metrics well-known in the literature. In the ACCEPT project, we will use a selection of these metrics including BLEU, GTM, Meteor and TER. The common basis of these metrics is to quantify the differences between a machine translation and one or more reference translations in terms of a mathematical formula. In this study, we will be assuming that reference translations are always post-edited translations. The metric we are ultimately most interested in is an estimate of the amount of effort required by the post-editing process to improve the initial machine translation to a level adequate for the task. Human evaluations will be used in conjunction with the automated scoring mechanism, the purely subjective balancing the purely objective, to guide the development of the machine translation technology.

GTM

GTM – *General Text Matcher* (Turian et al., 2003) – is a metric based on precision and recall. "Precision" measures how many of the words generated by the system are correct, while "recall" measures how many of the words that the system *should* generate are correct:

$$\text{precision} = \frac{\text{correct}}{\text{output-length}} \qquad \text{recall} = \frac{\text{correct}}{\text{reference-length}}$$

The "f-measure" combines precision and recall by considering their harmonic mean (precision and recall are both important – we do not want to output wrong words, but we do not want to miss correct words either):

$$\text{f-measure} = \frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})/2}$$

GTM computes precision, recall and f-measure in terms of maximum unigram matching, i.e., *correct* refers to the maximum subset of non-repeated words present in both output and reference. GTM favours longer matches and matches in the right order; these matches are assigned higher weights. The weight of longer matches is a parameter of the metric.

BLEU

BLEU – *A Bilingual Evaluation Understudy* (Papineni et al., 2001) – is currently the most popular automatic evaluation metric. It measures the similarity between a machine translation and a set of reference translations using matches at the level of *n-grams*. For each n, where n typically ranges from 1 to a maximum of 4, the *n-gram precision* is defined as the number of n-gram matches in the output divided by the total number of n-grams in the reference translation. The general BLUE-n formula score assigns each n-gram precision a given weight; however, in practice these weights are typically considered as uniform. A brevity penalty is considered for dropping words, so that the score is proportionally reduced if the output is too short. The BLEU-4 formula is:

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

Since the 4-gram precision is often 0, the BLEU score is commonly computed over the test set rather than on the sentence level. There are many refinements to the basic BLEU metric, for example Smoothed BLEU (Lin and Och 2004).

METEOR

METEOR – *Metric for Evaluation of Translation with Explicit ORdering* (Banerjee and Lavie, 2005; Denkowski and Lavie, 2011) – is an extension of the BLEU metric, and is based on both precision and recall. It puts more emphasis on the role of recall relative to precision, reflecting the intuition that recall is more important for ensuring that the output captures the meaning completely. To detect meaning equivalences conveyed by paraphrases, METEOR incorporates both matches at the stem level and synonym matches, whereas BLEU only considers matches between words with identical surface forms. The matching process is thus more computationally expensive. The METEOR formula involves many more parameters, such as the relative weight of recall to precision, and the weights for stemming and synonym matches. Like GTM, METEOR favours matches that are contiguous and identically ordered, and its formula includes a so-called *fragmentation penalty* with two parameters. All these parameters can be tuned to maximize correlation with human judgments.

TER

TER – *Translation Error Rate* (Agarwal and Lavie, 2008) – is an error metric which measures mismatches, not matches. Like WER (Word Error Rate, one of the first automatic evaluation metrics used in machine translation), TER is based on the Levenshtein distance, which counts the number of elementary operations – insertions, deletions and substitutions – needed to transform the output so that it matches the reference. TER adds to these an operation called *block movement* (also, *jump* or *shift*), allowing the movement of word sequences from one part of the output to another.

Analysis of metrics

Relative advantages and shortcomings

A great deal of activity in the MT field has been concerned with the design of automatic evaluation metrics. Right now, no automatic metric for MT quality seems to be mature enough to provide an intuitively sound evaluation score (Estrella, 2008). There is much debate on their value in comparing different types of systems and in reflecting human judgements. However, these methods are still useful during the development process; as (Koehn 2010) underlines, most MT researchers design their systems "based on the rise and fall of automatic evaluation scores".

Shallow vs. deep metrics. The main problem that affects all these metrics is that they rely on comparison with a limited set of reference sentences, and penalize perfectly acceptable translations that are different from the references. This fact accounts for the low scores assigned to human translations, as well as for the low correlation with human judgements observed in specific settings. When measuring the similarity between a translation and a reference sentence, most of the metrics focus on shallow levels, relying on tokenization and string identity (e.g., GTM, BLEU, TER), whereas ideally they should take into account deeper structural levels and account for paraphrases, i.e., different means of expressing the same meaning (METEOR and RTE[1] do this to a limited extent). The first group of metrics has the advantage of being easily applicable to many languages, whereas the second group is more computationally expensive and dependent on the availability of language-specific resources and tools. Moreover, the latter group may require parameter tuning when applied in a new evaluation context.

Precision-oriented metrics vs. precision-and-recall-oriented metrics. Metrics are supposed to capture the extent to which a translation preserves the meaning of the source sentence (*adequacy*), and therefore penalize a translation with missing words by integrating a score for recall. Among the metrics discussed, BLEU does not implement recall, but it compensates for this by including a brevity penalty. BLEU is centred on precision, which captures the conformity to language (or *fluency*). GTM and METEOR model both precision and recall, giving them either equal (GTM) or different (METEOR) weights. TER models recall as well, through the mechanism of insertion.

Order-dependent vs. order-independent metrics. Some metrics, such as GTM, METEOR and TER, are more order-dependent in that they favour matches of contiguous word sequences that appear in exactly the same order in the translation as in the reference. But the word order requirement they impose may be seen as too harsh. In view of the high morpho-syntactic variability of language, some credit should be given to near matches as well. In this respect, BLEU is less restrictive, and more

[1] (Pado et al., 2009).

order-independent. It captures both matches of unigrams in a position-independent way, and longer n-grams like order-dependent metrics.

Further criticisms apply to all the metrics discussed above:

- It is not clear what aspect of translation quality is measured by each metric.
- The metrics' score is meaningless; at least, nobody has a clear idea of what it means.
- The metrics ignore the relative importance of words (a frequently quoted example is the negation particle, *not*, which is given the same weight as a determiner such as *the*).
- The metrics operate at a local level, failing to address the overall grammatical coherence of the sentence.
- Some metrics are biased towards SMT systems (e.g., BLEU is biased towards phrase-based SMT systems).

Much effort has been devoted to addressing these issues and advancing the state of the art in automatic MT evaluation. The plan in the ACCEPT project is, however, to apply in each specialist area the metrics that are in common usage there. Human evaluations will be used in conjunction with the automated scoring mechanism; thus, the purely subjective will balance the purely objective to guide the development of our MT technology.

Suitability to the ACCEPT project

When applied to the scenario which forms the basis of the ACCEPT project, the main questions that arise regarding the application of automatic evaluation metrics are the following:

- Will the metrics help us accurately distinguish between RBMT and SMT systems? It is already known that automatic metrics are not fully suitable for comparing systems of different types in general, and rule-based with statistical systems in particular (Callison-Burch et al., 2006, p. 104; Koehn and Monz, 2006). In the ACCEPT project, we will however compare post-edited rather than raw translation output. In Task 9.4, we will assess the usefulness, adequacy and reliability of metrics in evaluating translated user-generated content (UGC), by determining their correlation with judgements provided by translators.
- Are the metrics suitable for evaluating translated UGC? The type of content considered in the ACCEPT project differs substantially from the news content typically used when evaluating MT systems. It is not clear how useful existing metrics are for dealing with this type of content, given that they rely on text pre-processing (tokenization, alignment, lexica); the presence of many unknown words as well as the lexical dispersion due to alternative spelling and inconsistent terminology definitely causes problems for pre-processing. Some metrics may behave more robustly than others in this respect. The effectiveness of metrics on UGC has been little addressed by existing research. Roturier and Bensadoun (2011) compared different MT systems on user forum data and found no consistent correlation between automatic scores and human comprehensibility judgements. They also found that the source quality impacts systems and language pairs differently, and that the choice of metrics has a significant impact on comparing systems (the rule-based system being heavily penalized by automatic metrics).
- A related question is whether the metrics are at all useful for comparing the output of the baseline MT systems against that of the enhanced MT systems developed in the project.

More specifically, the question that will be addressed in Task 9.1 is whether the metrics can accurately reflect the impact of pre-editing rules on translation quality. It is well known that metrics are insensitive to the relative semantic salience of words; for instance, a translation that preserves the semantic polarity of the source sentence is definitely preferable to a translation that does not, but will the difference in the automatic scores mirror this distinction?

- Given the particular task at hand – the design of an enhanced SMT system based on pre-editing – it is important to know whether automatic metrics are effective in locating potential areas where the system can be improved, i.e. finding which characteristics of the source text influence the translation quality most, so that linguists can design pre-editing rules accordingly. Preliminary experiments conducted in the framework of the ACCEPT project for the language pairs English-French and English-German concluded that metrics help identify specific cases where the precision of pre-editing rules could be improved (Roturier et al., 2012). The work in Task 9.1 will provide further insights on this issue.

Proposal for a task-oriented metric

The ACCEPT task differs significantly from ones previously considered in the literature; as pointed out, for example, by Doyon and her colleagues (1995), a good task-oriented metric needs to be adapted to the task under consideration. In our case, an appropriate metric will need to consider the following requirements:

- It should be appropriate to the specific evaluation context:
    - the **UGC domain**, in particular:
        - Symantec's Norton user forum content
        - Translation Without Borders' medical content
    - the **pre-editing** process (although our SMT system operates on UGC, the input is normalised in order to get rid of most errors)
    - the **post-editing** process (our SMT system will learn from post-editing in order to automate correction as much as possible)
    - (optionally) the enhancements made at the **SMT engine** level:
        - text analytics
        - sentiment analysis
        - domain adaptation
        - linguistic backoff.

  It is important to note that the evaluation of the impact of pre-editing and post-editing rules is in general less challenging than evaluation of large scale SMT systems, since the changes introduced are more local in nature.
- It should be applicable to **all language pairs** considered in the project, while aiming to be language independent:
    - English-French
    - English-German
    - English-Japanese
    - French-English.
- It should consider the ACCEPT experimental setting, which involves:

    - different types of **editors**:
        - monolingual vs. bilingual speakers
        - subject-matter experts vs. non-experts
        - Casual editors with no previous experience, e.g. ones recruited via the Amazon Mechanical Turk
    - different types of **output**:
        - raw translation output vs. post-edited output
    - the size of the **reference** set:
        - a single reference vs. multiple reference translations vs. no reference.
- It should integrate **multifaceted information** available elsewhere in the project, in particular:
    - in the **evaluation portal** built in WP5:
        - user ratings for source, translated and post-edited content
        - time spent editing;
        - reward scores (acting as credibility indicators)
    - in the source normalization process performed in WP2:
        - type, number and weight of pre-editing rules applied
        - the string distance between the two versions of the source content
        - the readability level attained.
- It should be able to quantify the impact of specific **pre-editing/post-editing rules** on translation quality, thus being useful in assessing the effectiveness of individual rules.
- It should **compare favourably** against the existing metrics considered in the project (GTM, BLEU, METEOR and TER) in terms of correlation with human judgments provided by end users, translators, and students.

The exact definition of the new metric will be the result of an experimental process that will test different ideas, by incorporating various types of information from the ones mentioned above as they become available. Thus, the finalisation of the metric definition and the implementation will occur at a later stage of the project. Right now, we consider that a rudimentary form of the new metric could be the linear combination of a subset of the variables considered, each one being assigned a given weight. For instance, one such variable could be the BLEU score, with a corresponding weight; another could be the average user rating, with the corresponding weight computed as a function of the users' reward scores.

An appealing feature would be the ability to perform evaluation without a reference translation. When no reference is available, the metric could back up to characteristics of the translation alone (a process similar to evaluation by monolingual speakers, who do not have knowledge of the source language). Thus, rather than relying on computation of sentence similarity, the metric could rely on the language model probability of the output and on its syntactic and semantic features, with the goal of measuring the well-formedness, fluency, and readability of the output considered in isolation. These requirements overlap to a considerable extent with those of the currently popular quality estimation task (http://www.statmt.org/wmt12/quality-estimation-task.html), and we will be paying careful attention to work in this area. In particular, we think that this method will be feasible for the most common case that arises when evaluating the impact of individual pre-editing rules: the effect of the rule is known when it fires, and the question is whether it has fired or not.

Bibliography

Agarwal, A. and Lavie, A. (2008). Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 115–118, Columbus, Ohio. Association for Computational Linguistics.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, pp. 85–91, Edinburgh, Scotland.

Doyon, J.B., White, J.S. and Taylor, K.B. (1995). Task-Based Evaluation for Machine Translation. MT Summit VII.

Estrella, P. (2008). *Evaluating Machine Translation in Context: Metrics and Tools*. PhD thesis, University of Geneva.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P. and Monz, Cr. (2006). *Manual and Automatic Evaluation of Machine Translation between European Languages*. In *Proc. of the Workshop on Statistical Machine Translation*, pp. 102-121.

Lin, C.Y. and Och, F.J. (2004). Orange: a method for evaluating automatic evaluation metrics for machine translation. Proc. COLING.

Pado, S., Galley, M., Jurafsky, D., and Manning, C. (2009). Robust Machine Translation Evaluation with Entailment Features. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp. 297–305, Suntec, Singapore.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Report.

Roturier, J., and Bensadoun, A. (2011). Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of MT Summit XIII: the Thirteenth Machine Translation Summit*, Xiamen, China.

Roturier, J., Mitchell, L., Grabowski, R., and Siegel, M. (2012). Using automatic machine translation metrics to analyze the impact of source reformulations. In *Proceedings of AMTA*, San Diego, CA, USA.

Turian, J., Shen, L., and Melamed, D. (2003). Evaluation of machine translation and its evaluation. In *Proceedings of the 9th Machine Translation Summit*, pp. 386–393, New Orleans, Louisiana.