Weakly supervised approaches for quality estimation

Erwan Moreau, Carl Vogel

Machine Translation manuscript No.
(will be inserted by the editor)

Weakly supervised approaches for Quality Estimation

Erwan Moreau · Carl Vogel

Received: date / Accepted: date

Abstract Currently, Quality Estimation (QE) is mostly addressed using supervised learning approaches. In this paper we show that unsupervised and weakly supervised approaches (using a small training set) perform almost as well as supervised ones, for a significantly lower cost. More generally, we study the various possible definitions, parameters, evaluation methods and approaches for QE, in order to show that there are multiple possible configurations for this task.

Keywords Quality Estimation · Evaluation · Unsupervised learning · Weakly supervised learning

This research is supported by Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) funding at Trinity College, University of Dublin.

E. Moreau
CNGL and Computational Linguistics Group
Centre for Computing and Language Studies
School of Computer Science and Statistics
Trinity College Dublin
Dublin 2, Ireland
Tel.: +353 1896 1885
E-mail: moreaue@cs.tcd.ie

C. Vogel
Computational Linguistics Group
Centre for Computing and Language Studies
School of Computer Science and Statistics
Trinity College Dublin
Dublin 2, Ireland
Tel.: +353 1896 1765
E-mail: vogel@cs.tcd.ie

1 Introduction

Quality Estimation (QE) aims to provide a quality indicator for machine translated sentences. There are many cases where such an indicator would be useful in a translation process: to compare different Machine Translation (MT) models on a given set of sentences, to tune automatically the parameters of a MT model, to select the bad sentences for human translation or post-editing, to select the good sentences for immediate publication and try to apply automatic post-editing to the others, or simply to provide users who are not fluent in the source language information about the fluency of the translated text they are reading. As long as machine translated text cannot be of reasonably consistent quality, QE is helpful in indicating linguistic quality variability.[1]

Current approaches in automated QE have naturally focused on supervised learning (for example in [2,15,4,14,1]), since it is usually the best way to obtain optimal results, especially in such a difficult task which depends on a variety of parameters. Nevertheless we think that it is also interesting to study alternative "weakly supervised" approaches, i.e. either unsupervised or supervised with a only small amount of examples. By definition, such methods cannot achieve as good results as the "fully supervised" approach, but they are less costly and more flexible. In fact, there is a trade-off between cost and performance, and our main focus in this paper is to show that taking this parameter into account is important in the case of QE.

From this perspective we will explore the range of intermediate settings which can be used, trying to emphasize the differences between those in terms of performance, cost, ease of use and applications. In particular, we will show that the gain in performance obtained in the fully supervised setting is not very high in general, whereas its cost in terms of data and human expertise is significant. We will start in §2 by presenting qualitatively the limitations of the fully supervised approach, and by contrast the advantages of the weakly supervised setting(s). In particular we think that there are numerous use-cases where the latter is more adequate than the former, for instance if resources or human expertise/time are missing. Moreover, one of our main focuses will be the ease of use and the flexibility that the weakly supervised case(s) offer.

In §3 we define formally the different settings that we study, as well as the data and methods we use in the experiments presented in the following sections. This brings up the question of evaluation, studied in §4: comparing in detail the different approaches is not trivial, because they correspond to different settings which can not (or not totally) be evaluated using the same tools; this is one of the reasons why we introduce a new evaluation measure, the Mean Relative Rank Error (MRRE): this is in fact only one of the various ways to use Relative Rank Error, a very simple and intuitive concept which permits comparison of performance in a more meaningful and flexible way than other metrics. Thus, this metric also serves our purpose of proposing QE tools which are easy to use and understand. More generally, we conduct

[1] We focus on translation fluency rather than target language faithfulness to sources.

a detailed comparison of the different evaluations methods, showing for each of them their advantages and weaknesses. In particular we observe that the Mean Average Error measure, used to evaluate QE as a scoring task and often also to train supervised models, can introduce a bias in the predicted scores.

Finally, we propose various sets of experiments to validate our assumptions. First, we test our evaluation methods in §5, which also provides some new insights on the results of the WMT12 Shared Task and more generally on the task of QE. Then we compare in §6 the fully supervised approach against the totally unsupervised one as ranking tasks, and prove that the difference in performance is rather low. Finally in §7 we detail how the supervised learning approach performs in the scoring task depending on the number of examples provided. We show that there is a range of possibilities between the unsupervised and supervised settings which brings flexibility and ease of use for a moderate cost. Overall, our aim in this paper is to broaden the scope of Quality Estimation in terms of possible definitions, approaches, evaluation methods and parameters. We think that the current choice in favor of supervised learning approaches in most works on QE results from the focus on performance; this is of course an important criterion, but it should not conceal the others.

2 Motivations

2.1 Different definitions of Quality Estimation

QE can be defined in several different ways, entailing different approaches:

- As a **scoring task**, the goal is to provide for every sentence a meaningful score which represents its quality. In the WMT12 Shared Task for instance, scores must belong to the range $[1, 5]$ and each value can be interpreted according to some precise predefined guidelines (see §3.1 below);
- As a **classification task**, the goal is to annotate every sentence with a label taken from a predefined set, e.g. $\{good, edit, bad\}$.
- As a **ranking task**, the goal is only to rank the set of sentences relatively to one another. These ranks are less interpretable than a quality score or a label, since they do not say if a sentence is good or bad–they only say if it is better or worst than each other one from the dataset.

It merits mention that it is non-trivial to obtain consistent human evaluation of linguistic quality, and that the choice of measurement can introduce artifacts. For example, a human ranking of sentences among each other for relative grammaticality on a a pairwise basis may yield a rank ordering that is not transitive; coarse grained categories (e.g. $\{good, edit, bad\}$) may not sufficiently characterize the range of variability in the data and yield unreliable classifications for items that are at the boundary of applicability for any label; the scoring task lacks face-validity since people do not appear to operate with a clearly graded grammaticality or acceptability assessment function [8].

Clearly the first two definitions are more helpful in the viewpoint of applications. In theory one can imagine using unsupervised techniques to carry these

tasks out (e.g. using unsupervised clustering, like K-means); however it would be very difficult to effectively predict quality this way, which is why most[2] previous works use supervised learning in order to provide reliable scores or labels. Thus, these tasks require preliminary training, usually using a set of sentences annotated individually with quality scores (or labels) from a predefined scale.

In contrast, the third definition does not (necessarily) need training: since there is no absolute scale to follow, it is possible to design a system which ranks sentences based only on the input data. For example, sentences can be ranked according to their length, assuming that longer sentences are more likely to contain errors. In such a case the process is totally unsupervised, but in this paper we extend the scope of what we call "unsupervised" to any system which does not require a training set *annotated with quality scores*; in particular, this does not exclude the use of other training material, which can typically be some plain text corpus used to train a language model (in this sense the method is actually supervised, but only weakly). Such a distinction makes sense here, since there are major differences between these two kinds of training data in terms of applications of QE (see §2.2 below). A ranking permits, for instance, extraction of the N% best or the N% worst sentences, in order to apply a different treatment to each such group of sentences. Usually a rank is determined by some kind of numeric value, but this underlying value is different from a quality score: it cannot be interpreted in a meaningful way in general. Ranking subsumes scoring; i.e., one can always generate a ranking based on a set of scores, whereas the converse does not hold.[3]

In the remaining of this study we do not consider the case of classification, which is less general that the case of scoring. We also only consider the case of simple unsupervised systems based on a single measure/feature (see §3.3).

2.2 Problems with supervised learning

State of the art quality estimation systems use supervised learning approaches [15,4,1]. Given a corpus of sentences annotated by experts with quality scores and a set of features, the system is trained using regression (or classification) algorithms, in order to make it able to predict quality scores on a new set of sentences. There are two main constraints in this approach:

- *Human expertise* is required to annotate the training set with quality scores. Building a good training set is not trivial: given that there can be a high level of disagreement between annotators, it is safest to have the same sentences annotated by more than one expert.[4] Often it remains necessary to

[2] Actually, to the best of our knowledge, all works which study QE in this setting use supervised learning.

[3] The transitivity problem mentioned above is averted by ranking sets of sentences in terms of their comparison to a separate reference corpus. A ranking system can be converted to a classification system by using thresholds, but in this case it usually requires some supervision. We do not study this case because we consider that the main advantage of QE as a ranking task is precisely the fact that it does not need supervision.

[4] Further, great care is necessary in setting the terms of the annotation exercise [8,9].

filter the training data, for example by making scores from different annotators more homogeneous or by discarding items with excessive disagreement (this occurred with the WMT12 Shared Task [1] data).
- *Lack of reusability* is caused by the fact that the target data must be similar to the training data, as in any supervised learning context. The model can only be applied to translations tasks with the same pair of source and target languages, but its performance will also depend on target sentences being close to the training data in terms of domain and style (the extent to which this similarity must hold depends a lot on the kind of features used, and possibly on the learning algorithm and the guidelines followed by the human annotators). Therefore, the human expertise required to build the training data will be needed for every distinct kind of document.

These constraints are acceptable if the results allow a significant improvement in final translation quality. However, currently the scores predicted by a state of the art QE system are not reliable: figure 1 shows how predicted and actual quality scores relate to each other for the best supervised learning system (SDLLW [12]) in the WMT12 QE Shared Task; a strong correlation is observed, but for a significant number of sentences, the predicted quality is far from actual quality. Furthermore, by definition, a quality estimation measure is unlikely to perform better than a (sentence level) MT evaluation metric[5] (e.g. HTER, Smooth BLEU, etc.), because the task consists of doing the same thing with less information. Consequently, even if state-of-the-art performance in QE improves in the future, it can probably never achieve high reliability.[6]

2.3 The interest of weakly supervised approaches

By definition, weakly supervised approaches cannot perform as well as supervised approaches. Their advantage, however, lies in:

- low cost in terms of data and human expertise, as opposed to supervised approaches (see §2.2 above).
- ease of use (unsupervised case only): typically an unsupervised system only requires some reference (monolingual) corpus, which is rather easy to obtain. Additionally, the system is much simpler than in the supervised case and can be used by non-experts.
- flexibility: with an unsupervised measure, the reference corpus can be chosen according to the domain/style of the sentences to annotate (it is easier to find an adequate plain text corpus rather than a suitable set of annotated sentences). The case of supervised learning is less flexible, but offers

[5] Assuming the evaluation metric is provided with enough reference sentences (see §5).

[6] In fact, if it were possible to obtain nearly perfect QE scoring, then nearly perfect machine translation would not be far, since an ideal MT system could be built in the following way: apply the initial MT system, then start again only for the output sentences which do not obtain high QE scores, and loop until every sentence obtains a high score (at each turn the system can be tuned in various ways to improve the quality, possibly using learning from the previous results, until convergence).

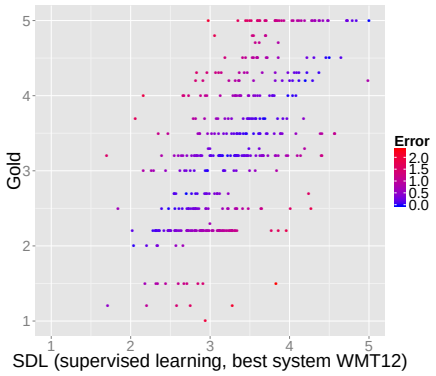

Fig. 1 Predicted score (X) vs. actual score (Y) for the best supervised learning system in WMT12 (SDLLW). A perfect system would be represented as the diagonal line $x = y$. Example: the red point at the top center represents a very good sentence (maximum score 5) which is predicted as mediocre (score 3).

the possibility to select the number of instances and/or features depending on the ratio quality/cost one wants to achieve (see §7).

There are a lot of applications where perfect quality is not required (sometimes not even expected), especially when the customers or users are mainly looking for low cost translations or products. For example, websites which provide a user forum can offer to their users the possibility to automatically translate a post. As long as the service is free, no one expects the translation to be of professional quality. In such a setting, QE does not have to be very reliable, and even a relative ranking by quality can be enough: it can be used to discard or translate manually the N% worst sentences, or to post-edit or publish immediately the N% best sentences. This is why unsupervised or weakly supervised approaches to QE can be relevant, especially when considering the fact that they can perform very well (see §6).

Finally we want to emphasize that the choice between supervised, weakly supervised or unsupervised QE, as well as the choice of other parameters of the system, depends on the parameters of the application one wants to achieve. In QE, as in many other tasks, there is a trade-off between cost and quality, and we will show in the next sections that a large range of strategies is available.

3 Definitions and description of the experimental framework

3.1 Definitions

When considering QE as a scoring task, a predefined scale is given which gives precise indications on how to interpret scores. In the next section we use the WMT12 Shared Task [1] dataset, which is based on the following guidelines:

1. The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.

2. About 50% -70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.
3. About 25-50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.
4. About 10-25% of the MT output needs to be edited. It is generally clear and intelligible.
5. The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation, but requires little to no editing.

When considering QE as a ranking task, the sentences are ordered from highest quality to lowest, without any indication of their absolute quality (the first sentence can actually be very bad or the worst still of high quality). Generally, a ranking is based on an underlying function which assigns a numerical value[7] to every sentence (e.g. the probability of the sentence in a given language model). A ranking assigns a *rank* to every sentence in the input set, usually w.r.t the order of the values sorted in the appropriate direction. Given a set of sentences $S = s_1, \ldots, s_n$, a ranking *without ties* is a bijection between S and the integers $1, \ldots, n$. Such a mapping satisfies the ranking property:

$$\sum_{s \in S} rank(s) = \sum_{1 \leq i \leq n} i \tag{1}$$

A ranking *taking ties into account* is also a function which assigns a rank $rank(s) \in [1, n]$ to any sentence s, but it is not necessarily a bijection and ranks are not always integers. We show in §4.1 that taking ties into account (cases where two sentences are considered of the same quality) can be important in a ranking task. In order to fulfill the ranking property, a set of tied sentences S are all assigned the average rank $r = \frac{r_{min} + r_{max}}{2}$, where r_{min} (resp. r_{max}) is the lowest (resp. highest) rank that any of these sentences could have been assigned in a ranking without ties. In order to make rankings comparable, the highest quality sentences are assigned the first (lowest) ranks by convention. Unless otherwise specified, we consider below only rankings with ties.

Spearman's correlation is the most standard way to compare two rankings. For any two sets of values, it is defined as Pearson's correlation applied to the rankings obtained by sorting these two sets (i.e., using Pearson's correlation on the ranking is the same as using Spearman's correlation on the values)[8].

3.2 Datasets

The experiments presented in the next sections use the data provided by the the WMT12 Shared Task on Quality Estimation [1]. This data consist in:

[7] We use the generic term *value* to prevent any confusion with the word *score*, used only to describe an absolute indication of quality according to a predefined scale.

[8] As a consequence, Spearman's correlation does not make any difference if the two sets of values come from different types of distributions.

- The training data is a set of 1,832 sentences translated from English to Spanish; the source and target sentences are provided, as well as a reference translation and a translation obtained by post-editing the output sentence; the quality scores are averaged over several human annotators, and have been post-processed carefully to avoid ambiguous cases [1].
- The test data is a set of 422 cases: source and target sentences, reference translations, post-edited translations and quality scores (only the source and target sentences were available to the participants during the task).
- Various sets of features (total: 3,026 features, possibly including duplicates); more precisely, for every set and every feature in this set, the value obtained by every sentence in both the training and test set; these feature sets are:
  - the set of "baseline features" provided by the task organizers;
  - the 21 different sets of features used by participants for both the training and test set, collected and released by the organizers after the task.
- Finally we also use the 21 actual submissions of the participants, i.e., the scores and ranks that they have computed for the test set.

Participants were allowed to submit two systems, and each system was allowed to compete in the ranking sub-task, the scoring sub-task or both. Of course, the fact that a training set was provided (and even a set of features) encouraged the participants to use supervised learning, and they all did.

3.3 Experimental setting

In addition to the 21 actual submitted systems using supervised learning (for which the ranks and scores are available), we use the individual features as unsupervised measures to generate rankings. This is achieved simply by sorting any numerical feature[9] by their values to obtain a ranking (possibly including ties). Since the goal is to generate multiple rankings of various quality levels, the features are converted blindly, without any filtering: every feature is taken into account, even if by definition it lacks face validity–since we do not know what the feature actually represents, in order to fulfill the convention that the top quality sentences are ranked first, it is sorted in both directions (ascending and descending order). Thus, every individual feature is converted to two distinct (opposite) rankings, hence we obtain 6,052 rankings based on individual features. Additionally, 6,112 language models trained on Europarl [5] have been computed with various combinations of parameters using SRILM[10] [16], and for every such model a ranking based on the probability of the sentences is generated. In order to evaluate the quality of all these predicted rankings and sets of scores, we use the scores annotated by human experts as gold standard (the gold ranks are obtained by sorting the scores).

[9] Boolean features are processed as numerical features with only 0 and 1 as values, thus providing interesting cases of rankings with lots of ties to study; if the feature contains undefined values, these are added at the end of the ranking.

[10] `http://www.speech.sri.com/projects/srilm/manpages/` – last verified: March 2013

3.4 Baselines

The classical method of averaging over several random rankings of sentences creates a possible baseline, providing absolute lower bound against which to evaluate plausible ranking methods. If ties are taken into account, another naive baseline system consists in not ranking anything, i.e., assigning the same rank $(n+1)/2$ to every sentence. This is similar to the random ranking, but it is more sophisticated in the sense that the middle rank is the optimal choice to minimise the rank error. A more realistic baseline derives from sentence length, which is a fairly good indicator of quality (the longer the sentence, the more likely it is to contain errors); Spearman's correlation of length against the gold ranking is 0.36 (for the source sentences in English; 0.37 for the target sentences in Spanish) for the WMT12 test set. As noted by Callison-Burch et al. [1], it is also interesting to compare the performance of rankings against an upper bound (especially for *DeltaAvg* values, see §4.2): for QE, upper bounds are supplied by sentence-level MT metrics, which compare the output sentence against the reference (gold standard) translation.

4 Comparison of different evaluation metrics for QE

4.1 The importance of taking ties into account

In the WMT12 Shared Task [1], ties were not allowed for the sake of simplicity; this is reasonable since the task was mainly designed in a supervised learning perspective, and applying regression techniques with multiple features is unlikely to yield many identical scores. However, ties are more likely in an unsupervised context, since the ranking is usually a direct function of a set of underlying values.[11] Furthermore, there are usually ties in the reference ranking, because human annotators tend to assign integer scores: for the 1832 sentences that the WMT12 training data contains, the annotators have assigned only 23 different values as quality scores.

Thus, not taking ties into account introduces some randomness in the results, because of the arbitrary order in which a ranking without ties sorts the sentences. Figure 2 shows that in most cases there is very little difference between taking ties into account or not: the two versions correlate strongly (Pearson's correlation is 0.995); however there are outliers, mostly when the correlation is low: the largest difference observed is 0.17. Even if large differences occur rarely, it is a theoretical error to compute Spearman's correlation on a ranking which possibly contains portions sorted arbitrarily. To avoid this, either one must compute several rankings in which tied sentences are sorted randomly and finally use the mean, or one must use a ranking with ties.

[11] Of course the unsupervised setting produces ties only if the values are discrete or if there are only a small number of distinct values. It also depends on the number of sentences. For instance, a ranking relying on the length of sentences is very likely to contain ties.

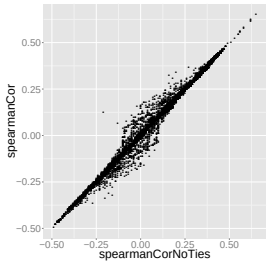

Fig. 2 Spearman correlation with ties (X) vs. without ties (Y) (cor.: 0.995).

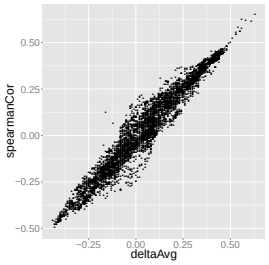

Fig. 3 DeltaAvg (X) vs. Spearman correlation with ties (Y) (cor.: 0.986).

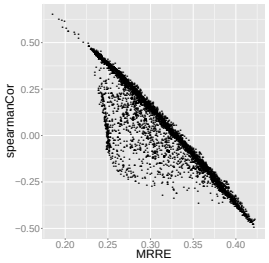

Fig. 4 MRRE (X) vs. Spearman correlation with ties (Y) (cor.: -0.946).

Relation between the different evaluation measures. Every predicted ranking has been compared to the gold ranking using different evaluation measures, and each point represents the result for a given ranking with two distinct measures (12,185 points). By construction some rankings are very bad (see §3.3), hence the negative values.

4.2 The Delta Average measure

One of the problems of the correlation coefficient is the difficulty to interpret it: it is commonly understood that the correlation is negligible under 0.2, low between 0.2 and 0.4, moderate between 0.4 and 0.6, etc.; but this does not say much about the performance for a QE indicator, and therefore does not help users to decide how much of the sentences they should post-edit (for instance). In the WMT12 Shared Task, the organizers introduced a new measure *DeltaAvg* [1], which does not actually compare two rankings together but a ranking against a reference set of scores. This requirement is usually easy to satisfy in an evaluation context (human annotators provide these scores); however it prevents comparing two (predicted) rankings together, and it also prevents comparing two *DeltaAvg* scores obtained on different datasets annotated with distinct scoring guidelines/scales. The goal is to make the resulting measure interpretable w.r.t the predefined scale used by the annotators in the following way: "for a given set of ranked entries, a value DeltaAvg of 0.5 means that, on average, the difference in quality between the top-ranked quantiles and the overall quality is 0.5" [1]. As a consequence, this measure depends on the distribution of the reference scores: if the evaluated sentences are of similar quality, the *DeltaAvg* score will be low even if the ranking is correct.[12]

$$DeltaAvg[n] = \frac{1}{n-1}\sum_{k=1}^{n-1} V(S_{1,k}) - V(S) \tag{2}$$

$$DeltaAvg = \frac{1}{N-1}\sum_{n=2}^{N} DeltaAvg[n] \tag{3}$$

[12] We have tested selecting randomly 200 cases from the WMT12 training set in two cases: among all scores and among scores between 2 and 3. In the first case the maximum *DeltaAvg* value (obtained by comparing the ranking to itself) is 0.86, but only 0.28 in the second case.

In these equations, $V(S)$ is the average quality score for a set of sentences S, $N = \frac{|S|}{2}$ and for a given number n of quantiles, $S_{i,j}$ is the union of the quantiles S_k for any k such that $i \leq k \leq j$.

Figure 3 shows that *DeltaAvg* correlates almost perfectly with Spearman's coefficient (correlation: 0.98). There can be quite large discrepancies (up to 0.16 between the *DeltaAvg* value predicted by linear regression and the actual one[13]), but these appear mostly around 0, whereas better QE methods reach higher scores. Some of these discrepancies might be due to the discrete nature of *DeltaAvg*: the definition states that if the number of sentences is not divisible by n when computing $DeltaAvg[n]$, the last quantile contains more sentences. This approximation might have an impact, especially on small datasets, but it could be fixed by averaging the quantiles using appropriate smoothing techniques. The other possible problem of *DeltaAvg* is its inability to take ties into account, which is also corrigible (in the same way as for the MRRE; see below). Finally, *DeltaAvg* is computationally intensive since all possible quantiles must be calculated, which might be an issue for large datasets.

4.3 Relative Rank Error (RRE) and Mean Relative Rank Error (MRRE)

We propose another simple measure to compare two rankings based on the difference between the predicted and actual rank. Let S be a set of sentences and *rank* a ranking on S. First the ranks are normalized in $[0, 1]$ w.r.t the size of the dataset: relative rank RR is defined as in (4) for any $s \in S$.[14]

$$RR(s) = \frac{rank(s) - 1}{|S| - 1} \tag{4}$$

Let RR_{gold} be the reference ranking; (5) defines Relative Rank Error (RRE).

$$RRE(s) = |RR(s) - RR_{gold}(s)| \tag{5}$$

The RRE is defined as the absolute value of the rank error because (1) it makes it more meaningful when averaged on a set of sentences and (2) the direction of the error does not matter in general: the low predicted ranks tend to be lower than their gold counterpart and conversely.

It is important to notice that, under this definition, the maximum possible value for RRE is not constant over the whole ranking: the maximum value 1 can be reached only at the extrema, i.e., if the best sentence is ranked as the worst or the opposite. Figure 4 shows that the mean of the RRE (MRRE) over the set of sentences is also strongly correlated to Spearman correlation, but less than *DeltaAvg* (Pearson's correlation is -0.95).[15] The differences visible in figure 4

[13] That is, if a regression line is drawn on figure 3, the maximum vertical distance between a point and the line is 0.16. While in this example the two measures take the same range of values, so that the regression line is close to the diagonal, this is not true in general.

[14] Both rank and size are decreased by 1 to rescale the ranks from $[1, n]$ to $[0, 1]$.

[15] The correlation is expected to be negative, since this metric measures error.

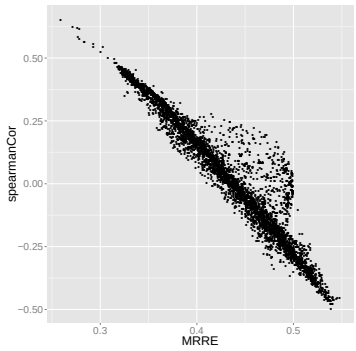

Fig. 5 Spearman correlation (Y) vs. alternative MRRE normalized locally (X) (cor.: -0.979). Compared to figure 4, the cases with lots of ties are penalised, which is why they appear on the other side of the regression line (WMT12, 12,000 points).

correspond to rankings containing many ties: in such cases, especially if a large majority of the sentences are assigned the same value, then by definition their average rank tends to move towards the middle of the ranking, which makes the final MRRE score lower; this is a consequence of the local variability of the maximum value. This issue can be solved by adopting the following alternative definition, which also normalizes w.r.t the local maximum error value (6).

$$RRE(s) = \frac{\mid RR(s) - RR_{gold}(s) \mid}{\max(\ RR(s)\ ,\ 1 - RR(s)\ )} \tag{6}$$

The correlation of this version of RRE with Spearman's correlation is stronger (-0.98), and the cases with many ties tend to be closer to the regression line (see figure 5). Nonetheless such cases are not frequent with real QE rankings. The alternative definition is not as intuitively meaningful as the first one, and makes the comparison of individual RRE scores less interpretable. This is why in the following we adopt the first definition, in particular because we aim to provide a measure which is understandable by non-experts.

The RRE score, which is defined at the sentence level, can be used to compute various statistics on the global ranking: mean (MRRE), median, extrema, standard deviation, etc. The MRRE is a simple case of Mean Absolute Error. The meaning of statistics based on RRE is straightforward: for instance, the MRRE says how far in average (w.r.t the size of the data) a sentence in this ranking is from where it should be. This is a useful information for applications of QE, especially because it can be applied to a subset of the ranking.[16] Thus, it is possible to compare rankings not only by their global similarity against the gold ranking, but also by their local performance. Since often the user is only interested in selecting the bottom or top N% of the ranking, it is useful to be able to compare ranking methods only on these parts of the ranking.

Ties require a special treatment when computing the MRRE on a subset of sentences, since some of the ties may be with sentences outside the

[16] *DeltaAvg* could probably be adapted to allow this possibility as well.

subset. The RRE values are smoothed for all sequences of tied ranks by assigning to each of these sentences the mean of the RRE values in this sequence. For example, let $(A,1)(B,2)(C,3.5)(D,3.5)(E,5)(F,6)$ be the gold ranking and $(B,1),(D,2),(A,4),(E,4),(F,4),(C,6)$ the predicted ranking; the RRE values are $(B,0.2),(D,0.3),(A,0.6),(E,0.2),(F,0.4),(C,0.5)$. Then the smoothed RRE values are $(B,0.2),(D,0.3),(A,0.4),(E,0.4),(F,0.4),(C,0.5)$, because A,E and F are tied and their average RRE is 0.4. Thus the MRRE for the bottom 50% of the predicted ranking is $(0.4+0.4+0.5)/3 = 0.43$. This prevents arbitrarily selecting the values of the sentences inside the subset (here E and F), since the tied sentences outside the subset (here A) are equally valid candidates.[17]

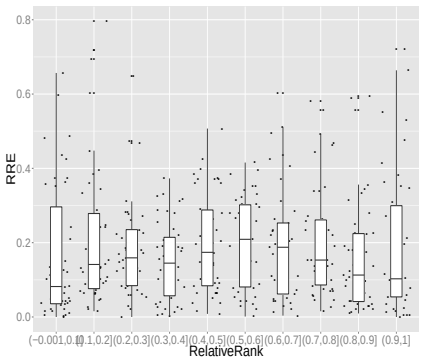

Fig. 6 RRE by sentence sorted by predicted (relative) rank (precision).

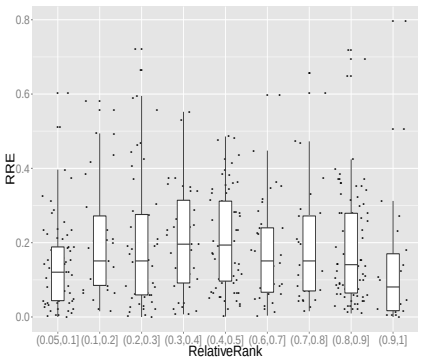

Fig. 7 RRE by sentence sorted by gold (relative) rank (recall).

Local RRE by 10% ranges. Each boxplot represents the RRE values for the subset of sentences which belong to the range N to $N+10\%$ of the predicted ranks (left side) or the gold ranks (right side). Data: SDDLW, best ranking at WMT12.

Additionally RRE can be used in various ways to provide helpful visual information about the ranking. For example figure 6 shows the RRE obtained by the ranking which performed best in the Shared Task (SDLLW, [12]), sorted by the relative rank predicted by the system. This shows the amount of error in different parts of the ranking. The X axis can also be sorted by the gold relative rank, as in figure 7. This shows a different kind of information, similar to a recall measure: one can see what parts of the actual ranking the system fails to rank correctly. This is useful for example if the user wants to extract the "bad" sentences and publish the "good" ones as they are: in this case it is more important to catch the maximum number of bad sentences (recall) rather than ensuring that there are no "good" sentences among the extracted part. For instance, figure 7 shows that among the 10% worst sentences (range $[0.9,1]$ on the X axis), 75% (third quartile) are predicted within a RRE of approximately 0.175 (i.e., 75% of the 10% worst sentence are never ranked farther than 17.5%

[17] However this can introduce a bias in the statistics based on quantiles, e.g. in the previous example it is not possible to decide what the maximum value of the bottom 50% is, because A can be considered either inside or outside the interval since it is tied to E and F.

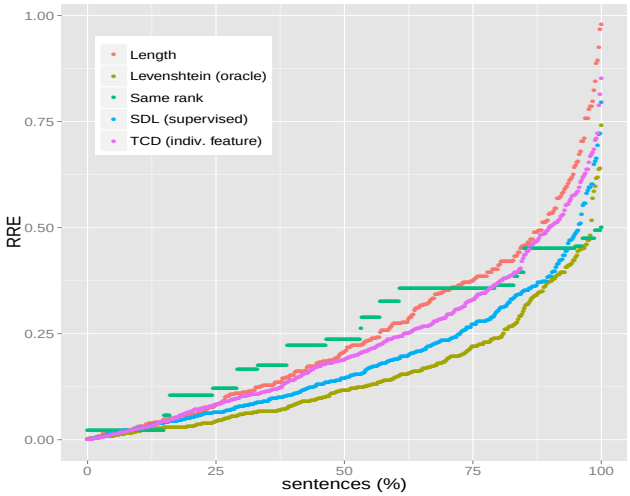

Fig. 8 Sentences sorted by RRE: the points show how many sentences (X axis) obtain an error level lower or equal to the RRE (Y axis) (WMT12 test data). Example: the SDL ranking assigns a rank within +/-26% w.r.t the gold rank for 75% of the sentences. The rankings represented here are: the length baseline (cor. 0.36, MRRE 0.25), the "all ranks equal" baseline (MRRE 0.25), the best individual feature (cor. 0.50, MRRE 0.23), the best supervised system (cor. 0.65, MRRE 0.19), and the Levenshtein edit distance (oracle based on the postedited sentence; cor. 0.74, MRRE 0.15). The plateaus observed for the "all ranks equal" curve are due to the ties in the gold standard ranking (more visible because this ranking contains only ties). A perfect ranking would correspond to the line y=0 (RRE=0 for every sentence).

from their actual position). In both graphs the total amount of error is the same, since the RRE is symetrical for the predicted and gold ranking.

Finally, figure 8 shows another interesting application of RRE: here the sentences are sorted by RRE, so that the plotted points show how many sentences (X axis) obtain an error level lower or equal to the RRE (Y axis). In other words, each point (X,Y) on this graph can be read "X% of the sentences are predicted within an Y% rank error margin". This kind of information is easily understandable by a non-expert user, and is also a valuable statistic for informing choice of a threshold to balance between quality and cost.

4.4 Evaluation for the scoring task: Mean Absolute Error

QE can be evaluated as a scoring task with standard measures like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), but here we use only the former, the primary measure in the WMT12 Shared Task [1].

The main advantage of the scoring setting over the ranking setting is to provide an absolute quality score, whereas the latter can only provide a relative indication of quality between sentences. But it seems that the supervised learning approach introduces a bias in the reliability of predicted scores as

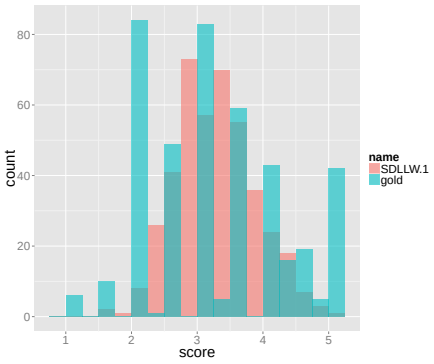

Fig. 9 Distribution of predicted scores vs. gold scores (WMT12 test set). Here the predicted scores are taken from the best system, SDLLW.1. The spread of the predicted scores is noticeably smaller than the spread of the gold scores.

absolute values. In the 19 sets of predicted scores submitted by the participants to the Shared Task (two outliers are discarded: DFKI_Maja and TCD.1), scores spread significantly less than the actual scores: the standard deviation (by group of scores) ranges from 0.33 to 0.94, with a mean of 0.51 and median of 0.47, whereas the actual standard deviation (for gold scores) is 0.98. This is probably caused by the fact that most supervised learning algorithms optimize the model by minimising the MAE, which favors assigning scores in the middle of the interval rather than close to the extrema. Additionally, the means range from 2.97 to 3.63, with a mean at 3.39 and a median at 3.43, whereas the mean of the actual scores is 3.29. This is probably due to the fact that the systems had been trained on a dataset for which the mean of the scores was 3.45. Figure 9 shows an example of the difference between the distribution of predicted scores and real scores.

It would be necessary to study more datasets and supervised systems in order to confirm this observation. If the conclusion holds, this is a problem. Since the interest of QE as a scoring task is precisely to provide an absolute scale of the quality level, if sentences at quality extremes tend to be predicted as moderately good/bad then the QE system lacks construct validity.

5 Various observations on the WMT12 results using MRRE

The global results of the 6 best systems in the WMT12 Shared Task are given in table 1, including RRE-based scores. Despite the overall consistency of the different metrics (which is due to the fact that the scores are high, as shown in §4), it is interesting to notice that a few differences appear. The UEDIN ranking is apparently negatively affected by the DeltaAvg scoring, since its other results tend to show that it is better than the TCD one. The SDL ranking is clearly above all the others for all the global evaluation metrics, but its median RRE is very close to the one obtained by UEDIN: both systems

System	DeltaAvg	Spearman no ties	Spearman	Mean RRE	Median RRE	Std. dev. RRE	Max. RRE
Lev. 2	*0.78*	*0.74*	*0.74*	*0.153*	*0.118*	*0.139*	*0.741*
S-BLEU 2	*0.74*	*0.69*	*0.69*	*0.168*	*0.125*	*0.152*	*0.812*
SDLLW.1	0.63	0.65	0.65	0.185	0.145	0.153	0.796
SDLLW.2	0.61	0.62	0.62	0.199	0.162	0.155	0.836
UPPSALA.2	0.58	0.62	0.62	0.196	0.151	0.158	0.728
Lev. 1	*0.58*	*0.55*	*0.56*	*0.208*	*0.151*	*0.176*	*0.897*
UPPSALA.1	0.56	0.62	0.62	0.193	0.160	0.159	0.806
TCD.2	0.56	0.58	0.58	0.202	0.151	0.173	0.878
S-BLEU 1	*0.56*	*0.50*	*0.51*	*0.222*	*0.172*	*0.181*	*0.894*
UEDIN.1	0.54	0.58	0.59	0.197	0.146	0.174	0.956

Table 1 Performance of the 6 best systems in the WMT12 Shared Task. The lines in italic correspond to "oracle" rankings: they are based on the comparison between the output sentence and a reference translation (1) or the postedited translation of the sentence (2) (*Lev.* is the Levenshtein edit distance on words, *S-BLEU* is the smooth BLEU score computed with unigrams to 4-grams).

are able to rank correctly most of the sentences with less than 15% rank error. Finally the system which best avoids very large errors is UPPSALA.2, which never positions a sentence more than 73% of ranks from where it should be.

The "oracle" rankings based on post-edited versions of the sentences outperform the other rankings as expected, however the ones based on the reference translation fall in the bottom of the ranking. If Spearman's correlation or the MRRE are considered instead of *DeltaAvg*, they are even quite far from these rankings. This raises two questions:

- From the evaluation viewpoint, this large difference between *DeltaAvg* and the other metrics is surprising, since it has been observed in §4 that the measures tend to agree closely when the correlation is strong (high values for *DeltaAvg* or Spearman, low values for MRRE). This shows that this is not always true, and it probably depends somehow on the way the underlying scores are built (since this is the main difference in these cases).
- More generally, it is difficult to interpret the fact that "oracle" rankings based on reference sentences do not perform as well as QE rankings trained on quality scores. While MT metrics based on a single sentence do not correlate strongly with human judgement (particularly when different lexical choices are possible), the quality annotations in the training set might carry more information than a reference translation of the sentence.

Figure 10 shows the (MRRE) performance of the 6 best systems in the top/bottom 10/20% of the ranking. As mentioned (§2), we are especially interested in differences in these subsets of the rankings because they are crucial parts of numerous applications of QE. Indeed, the different systems do not obtain the same results, which means that the systems do not catch exactly the same sentences in these parts of the ranking (this would have been the case if there were very good/bad sentences which were easy to recognize by a supervised system). In particular, a user would not always choose the SDL system depending on the kind of task he or she intends to do (for example, the SDL system is the worst in precision in the bottom 20%). One can also

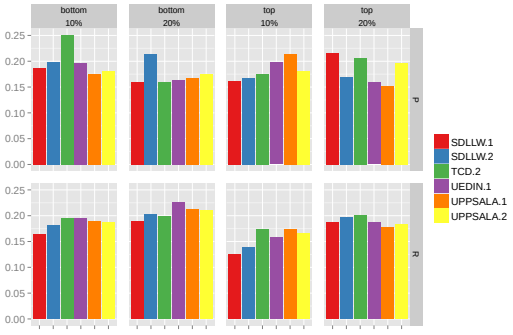

Fig. 10 Detailed MRRE in the top/bottom of the ranking for the 6 best systems at WMT12; P = Precision (subset of the predicted ranking), R = Recall (subset of the gold ranking) (test set, lower is better).

observe that most scores in these subsets are lower than the global MRRE, which ranges from 0.185 to 0.202 for these systems: the mean of the F1-scores for all the cases (top and bottom, 10 and 20%) is 0.183.

The fact that different systems catch different sentences is also confirmed by the correlation between rankings, which is not high in general: systems created by the same team often have a high correlation (between 0.8 and 0.95) together, otherwise the correlation coefficents fall between 0.7 and 0.9 (6 best systems), and the mean correlation is only 0.8 (6 best) or 0.56 (all systems) (see table 2). This could mean that there is still room for improvement for supervised learning approaches by combining the features of different systems.

Set of pairs of rankings	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6 best systems	0.7362	0.7744	0.8157	0.8206	0.8613	0.9364
all 21 systems	0.05097	0.31570	0.61180	0.55930	0.78200	0.95930

Table 2 Correlation coefficient statistics between pairs of WMT12 submitted rankings

6 The unsupervised setting using individual features

Table 3 and 4 show the performance of the best individual features on the WMT12 data, respectively on the training set and the test set (the latter was easier to rank than the former). The rankings based on these individual features make more errors than the supervised learning models presented above. Nevertheless, given the disadvantage that the lack of training represents, these values are not bad; in particular, the MRRE is around 0.25 (training data) and around 0.23 (test data) when the best scores obtained by supervised learning are between 0.18 and 0.20. In other words, the rank error is decreased by only a few percentage points (around 5%) on average when using a training

dataset and multiple features. For applications, such a small drop in the error rate might not be worth the alternative cost of supervision. The oracle method based on S-BLEU score on the reference sentence (0.222) is only 0.005 points better than the best individual feature on the test set (0.227).

Id	Del.Avg	Spear.	MRRE
dfki-el 65	0.41	0.397	0.246
uppsala 41	0.40	0.397	0.248
uppsala 42	0.40	0.397	0.248
sdl 5	0.40	0.395	0.249
dcu-symc 20	0.40	0.394	0.250
dfki-el 151	0.40	0.394	0.250
sdl 10	0.40	0.394	0.250
wlv-shef 53	0.40	0.392	0.250
dcu-symc 295	0.40	0.392	0.250
upc 44	0.40	0.392	0.250

Table 3 Best individual features, train set

Id	Del.Avg	Spear.	MRRE
tcd 28	0.49	0.496	0.227
uppsala 41	0.48	0.479	0.228
uppsala 42	0.48	0.479	0.228
dcu-symc 20	0.48	0.479	0.228
dfki-el 151	0.48	0.479	0.228
sdl 10	0.48	0.479	0.228
dcu-symc 34	0.48	0.462	0.231
dcu-symc 29	0.48	0.462	0.231
dcu-symc 39	0.48	0.461	0.231
phhlt 80	0.47	0.468	0.231

Table 4 Best individual features, test set

Contrary to the supervised case, the correlation between these best rankings based on individual features is very high: comparing any pair of rankings among the 10 best for the WMT12 training set gives a correlation between 0.86 and 1.0, with a mean of 0.96. This is expected since most of these features are based on language modeling. This is probably the main weakness in this kind of unsupervised approach: being a very basic method, it can not (or not easily) be improved by combinining multiple features as in the supervised case.

7 Supervised setting with a small amount of examples

In this section we study the setting of supervised learning using a limited number of examples as training data. Since training data is available in this framework, the system can predict absolute scores instead of only relative ranks. For applications, the goal is to minimise the number of instances needed while maximising the performance, depending on the context of use (thus offering a lot of flexibility). In this study, this is also an intermediate configuration which fills the gap between the unsupervised approach and the fully supervised (with many examples) in terms of evaluation. Since supervised learning can use any number of features and instances, it covers the whole range of settings, thus offering a more consistent and meaningful way to interpret evaluation results. Nevertheless the present study is only a short overview: our focus is on comparing the different cases, not on absolute performance; moreover only a few cases are studied, with arbitrary selections of features, and not using attributes selection techniques or other relevant methods. The experiments in this section used Weka [3], with linear regression, the SMO algorithm for SVM regression [11,10] and the M5P algorithm for regression with decision trees [6,17].

Figure 11 shows how supervised learning performs depending on the number of instances used as training data in various cases: the 5 first cases correspond to single features, and the 4 last cases to arbitrary (but good) selections

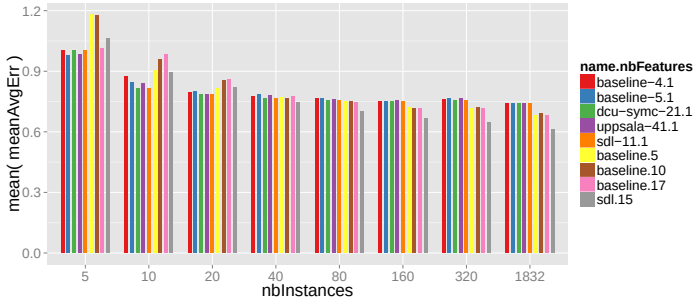

Fig. 11 Performance of supervised learning depending on the training data size (WMT12). Mean Average Error (lower is better) as a function of the number of instances in the training set (M5P, showing the mean of 10 random selections).

of multiple features. 10 random selections of sentences from the WMT12 training set were used as training data in each case, and the mean of their Mean Absolute Error (MAE) scores is shown. For simplicity, only the case of the M5P regression algorithm is shown; SVM regression gives similar results, and linear regression makes single features converge a little faster but does not perform as well for multiple features. We follow [13] in the idea that decision trees have the advantage of transparent interpretation, thus providing useful information about the main strengths and weaknesses in the translated sentences.

As expected, the MAE decreases with more instances very quickly for single features, since these cases do not need much training. The MAE is almost stable with only 20 instances; for multiple features convergence is slower, especially for the best set of features (SDL); naturally the cost of improving performance increases with the number of instances. For instance, it requires only 80 new instances to lower the MAE from 0.70 (80 instances) to 0.66 (160) whereas it requires 320 instances to reach 0.63 (640 instances) from 0.64 (320).

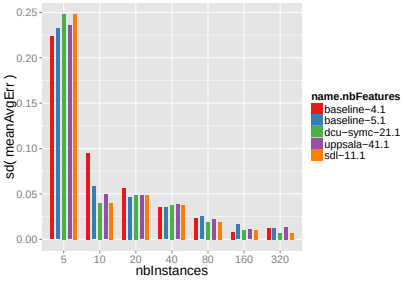

Fig. 12 Standard deviation for single features, M5P.

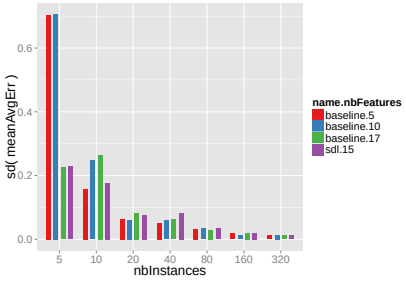

Fig. 13 Standard deviation for multiple features, M5P.

Since this setting relies on selecting a random subset of training set instances, it is important to study how this selection impacts the performance.

Figure 12 and 13 show the standard deviation observed among the 10 random selections on the MAE score. Unsurprisingly, the variation with single features is lower than with multiple features (which need more instances to converge), and it decreases as more instances are provided: e.g., the variation is lower than 0.05 with only 20 instances in the first case, 80 instances in the second.

Finally, it makes sense to pick the training data directly from the test set.[18] On the WMT12 dataset, using exactly the same settings as above but only for the cases 5,10,20,40,80,160 instances, training the model on instances from the test set instead of the training set decreases MAE on average by 0.017.[19]

These results show that weakly supervised learning can offer decent performance with drastically lower costs (number of annotated sentences) than the fully supervised framework. It does not suffer from the lack of an absolute scale (scoring task) as the unsupervised approach does. Moreover, in QE applications, one can select the appropriate number of features and instances depending on the target quality, the cost (or available human annotators/time) and the target level of confidence (typically based on the expected standard deviation), thus offering more flexiblity in the quality/cost trade-off.

8 Conclusion and future work

This paper has shown that there are reasons to study QE not only from the angle of performance. In particular, there are less costly possibilities than learning a model from a large set of annotated sentences which perform decently. We have also explored various ways to evaluate this task, including MRRE which is a very intuitive measure and can provide detail about the most crucial subsets of sentences.

Next steps include study of the framework of active learning for QE. We think that this might be an interesting alternative to using a large training dataset, because if there are not too many sentences these can be annotated directly by the user. This avoids the problem of the domain of the training set and gives the user more control over the quality criteria.

Acknowledgements We are grateful to the organizers of the WMT12 Shared Task on Quality Estimation for not only organizing the task but also collecting and releasing the sets of features of all participants afterwards. We also thank the reviewers for their valuable comments.

This research is supported by Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`www.cngl.ie`) funding at Trinity College, University of Dublin.

[18] Especially for QE, since in applications it is unlikely that all users can find easily a suitable training set from the same domain as the one they want to assess. Furthermore, the user is more likely to know their data and target result quality, in which case they can choose criteria to define the quality score scale and obtain results closer to what they expect.

[19] This improvement is statistically significant: a paired Student T-test gives a p-value of 0.003 (the population is 1620: 3 algorithms $\times$ 10 random selections $\times$ 6 sizes $\times$ 9 cases). In order to make a fair comparison, the selected instances are assigned their predicted value, not their human-annotated score (this would give better results in real applications).

Most calculations were performed on the Lonsdale cluster maintained by the Trinity Centre for High Performance Computing. This cluster was funded through grants from Science Foundation Ireland. The graphics in this paper were created with `R` [7], using the `ggplot2` library [18].

References

1. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2012 workshop on statistical machine translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation. Association for Computational Linguistics, Montreal, Canada (2012)
2. Gamon, M., Aue, A., Smets, M.: Sentence-level MT evaluation without reference translations: Beyond language modeling. In: Proceedings of EAMT, pp. 103–111 (2005)
3. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
4. He, Y., Ma, Y., van Genabith, J., Way, A.: Bridging SMT and TM with translation recommendation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 622–630. Association for Computational Linguistics (2010)
5. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit, vol. 5 (2005)
6. Quinlan, J.: Learning with continuous classes. In: Proceedings of the 5th Australian joint Conference on Artificial Intelligence, pp. 343–348. Singapore (1992)
7. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2012). ISBN 3-900051-07-0
8. Schütze, C.: The Empirical Base of Linguistics: Grammaticality Judgements and Linguistic Methodology. University of Chicago Press (1996)
9. Schütze, C.: Thinking about what we are asking speakers to do. In: S. Kepsar, M. Reis (eds.) Linguistic Evidence: Empirical, Theoretical and Computational Perspectives, pp. 457–485. Mouton De Gruyter (2005). Studies in Generative Grammar 85
10. Shevade, S., Keerthi, S., Bhattacharyya, C., Murthy, K.: Improvements to the SMO algorithm for SVM regression. Neural Networks, IEEE Transactions on **11**(5), 1188–1193 (2000)
11. Smola, A., Schölkopf, B.: A tutorial on support vector regression. Statistics and computing **14**(3), 199–222 (2004)
12. Soricut, R., Bach, N., Wang, Z.: The SDL Language Weaver systems in the WMT12 Quality Estimation shared task. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 145–151. Association for Computational Linguistics, Montréal, Canada (2012). URL `http://www.aclweb.org/anthology/W12-3118`
13. Soricut, R., Echihabi, A.: Trustrank: Inducing trust in automatic translations via ranking. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 612–621. Association for Computational Linguistics (2010)
14. Specia, L., Hajlaoui, N., Hallett, C., Aziz, W.: Predicting machine translation adequacy. In: Machine Translation Summit XIII. Xiamen, China (2011)
15. Specia, L., Turchi, M., Cancedda, N., Dymetman, M., Cristianini, N.: Estimating the sentence-level quality of machine translation systems. In: Proceedings of the 13th Conference of the European Association for Machine Translation, pp. 28–35 (2009)
16. Stolcke, A., et al.: SRILM-an extensible language modeling toolkit. In: Proceedings of the international conference on spoken language processing, vol. 2, pp. 901–904 (2002)
17. Wang, Y., Witten, I.: Induction of model trees for predicting continuous classes (1996)
18. Wickham, H.: ggplot2: elegant graphics for data analysis. Springer New York (2009)