UAlacant word-level machine translation quality estimation system at WMT 2015

Miquel Esplà-Gomis Felipe Sánchez-Martínez Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Abstract

This paper describes the Universitat d'Alacant submissions (labelled as UAlacant) for the machine translation quality estimation (MTQE) shared task in WMT 2015, where we participated in the word-level MTQE sub-task. The method we used to produce our submissions uses external sources of bilingual information as a *black box* to spot sub-segment correspondences between a source segment S and the translation hypothesis T produced by a machine translation system. This is done by segmenting both S and T into overlapping sub-segments of variable length and translating them in both translation directions, using the available sources of bilingual information *on the fly*. For our submissions, two sources of bilingual information were used: machine translation (Apertium and Google Translate) and the bilingual concordancer Reverso Context. After obtaining the sub-segment correspondences, a collection of features is extracted from them, which are then used by a binary classifer to obtain the final "GOOD" or "BAD" word-level quality labels. We prepared two submissions for this year's edition of WMT 2015: one using the features produced by our system, and one combining them with the baseline features published by the organisers of the task, which were ranked third and first for the sub-task, respectively.

1 Introduction

Machine translation (MT) post-editing is nowadays an indispensable step that allows to use machine translation for dissemination. Consequently, MT quality estimation (MTQE) (Blatz et al., 2004; Specia et al., 2010; Specia and Soricut, 2013) has emerged as a mean to minimise the post-editing effort by developing techniques that allow to estimate the quality of the translation hypotheses produced by an MT system. In order to boost the scientific efforts on this problem, the WMT 2015 MTQE shared task proposes three tasks that allow to compare different approaches at three different levels: segment-level (sub-task 1), word-level (sub-task 2), and document-level (sub-task 3).

Our submissions tackle the word-level MTQE sub-task, which proposes a framework for evaluating and comparing different approaches. This year, the sub-task used a dataset obtained by translating segments in English into Spanish using MT. The task consists in identifying which words in the translation hypothesis had to be post-edited and which of them had to be kept unedited by applying the labels "BAD" and "GOOD", respectively. In this paper we describe the approach behind the two submissions of the Universitat d'Alacant team to this sub-task. For our submissions we applied the approach proposed by Esplà-Gomis et al. (2015b), who use black-box bilingual resources from the Internet for word-level MTQE. In particular, we combined two on-line MT systems, Apertium[1] and Google Translate,[2] and the bilingual concordancer Reverso Context[3] to spot sub-segment correspondences between a sentence S in the source language (SL) and a given translation hypothesis T in the target language (TL). To do so, both S and T are segmented into all possible overlapping sub-

[1] http://www.apertium.org
[2] http://translate.google.com
[3] http://context.reverso.net/translation/

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 309–315,
Lisboa, Portugal, 17-18 September 2015.

segments up to a certain length and translated into the TL and the SL, respectively, by means of the sources of bilingual information mentioned above. These sub-segment correspondences are used to extract a collection of features that is then used by a binary classifier to determine the final word-level MTQE labels.

One of the novelties of the task this year is that the organisation provided a collection of baseline features for the dataset published. Therefore, we submitted two systems: one using only the features defined by Esplà-Gomis et al. (2015b), and another combining them with the baseline features published by the organisers of the shared task. The results obtained by our submissions were ranked third and first, respectively.

The rest of the paper is organised as follows. Section 2 describes the approach used to produce our submissions. Section 3 describes the experimental setting and the results obtained. The paper ends with some concluding remarks.

2 Sources of bilingual information for word-level MTQE

The approach proposed by Esplà-Gomis et al. (2015b), which is the one we have followed in our submissions for the MTQE shared task in WMT 2015, uses binary classification based on a collection of features computed for each word by using available sources of bilingual information. These sources of bilingual information are obtained from on-line tools and are used on-the-fly to detect relations between the original SL segment S and a given translation hypothesis T in the TL. This method has been previously used by the authors in other cross-lingual NLP tasks, such as word-keeping recommendation (Esplà-Gomis et al., 2015a) or cross-lingual textual entailment (Esplà-Gomis et al., 2012), and consists of the following steps: first, all the overlapping sub-segments σ of S up to given length L are obtained and translated into the TL using the sources of bilingual information available. The same process is carried out for all the overlapping sub-segments τ of T, which are translated into the SL. The resulting collections of sub-segment translations $M_{S\rightarrow T}$ and $M_{T\rightarrow S}$ are then used to spot sub-segment correspondences between T and S. In this section we describe a collection of features designed to identify these relations for their exploitation for word-level MTQE.

2.1 Positive features

Given a collection of sub-segment translations $M = \{\sigma, \tau\}$, such as the collections $M_{S\rightarrow T}$ and $M_{T\rightarrow S}$) described above, one of the most obvious features consists in computing the amount of sub-segment translations $(\sigma, \tau) \in M$ that confirm that word t_j in T should be kept in the translation of S. We consider that a sub-segment translation (σ, τ) confirms t_j if σ is a sub-segment of S, and τ is a sub-segment of T that covers position j. Based on this idea, we propose the collection of positive features Pos_n:

$$\mathrm{Pos}_n(j, S, T, M) = \frac{|\{\tau : (\sigma, \tau) \in \mathrm{conf}_n(j, S, T, M)\}|}{|\{\tau : \tau \in \mathrm{seg}_n(T) \wedge j \in \mathrm{span}(\tau, T)\}|}$$

where $\mathrm{seg}_n(X)$ represents the set of all possible n-word sub-segments of segment X and function $\mathrm{span}(\tau, T)$ returns the set of word positions spanned by the sub-segment τ in the segment T.[4] Function $\mathrm{conf}_n(j, S, T, M)$ returns the collection of sub-segment pairs (σ, τ) that confirm a given word t_j, and is defined as:

$$\mathrm{conf}_n(j, S, T, M) = \{(\sigma, \tau) \in M : \tau \in \mathrm{seg}_n(T) \wedge \sigma \in \mathrm{seg}_*(S) \wedge j \in \mathrm{span}(\tau, T)\}$$

where $\mathrm{seg}_*(X)$ is similar to $\mathrm{seg}_n(X)$ but without length constraints.[5]

We illustrate this collection of features with an example. Suppose the Catalan segment S ="Associació Europea per a la Traducció Automàtica", an English translation hypothesis T ="European Association for the Automatic Translation", and the most adequate (reference) translation T'="European Association for Machine Translation". According to the reference, the words *the* and *Automatic* in the translation hypothesis should be marked as BAD: *the* should be removed and *Automatic* should be replaced by *Machine*. Finally, suppose that the collection $M_{S\rightarrow T}$ of sub-segment pairs (σ, τ) is obtained by applying the available sources of bilingual information to translate into English the sub-segments in S up to length 3:[6]

[4] Note that a sub-segment τ may be found more than once in segment T: function $\mathrm{span}(\tau, T)$ returns all the possible positions spanned.

[5] Esplà-Gomis et al. (2015b) conclude that constraining only the length of τ leads to better results than constraining both σ and τ.

[6] The other translation direction is omitted for simplicity.

$M_{S\rightarrow T}$ ={***("Associació", "Association")***, ***("Europea", "European")***, ***("per", "for")***, *("a", "to")*, ***("la", "the")***, ***("Traducció", "Translation")***, ***("Automàtica", "Automatic")***, ***("Associació Europea", "European Association")***, *("Europea per", "European for")*, ***("per a", "for")***, *("a la", "to the")*, *("la Traducció", "the Translation")*, *("Traducció Automàtica", "Machine Translation")*, ***("Associació Europea per", "European Association for")***, *("Europea per a", "European for the")*, ***("per a la", " for the")***, *("a la Traducció", "to the Translation")*, *("la Traducció Automàtica", "the Machine Translation")*}

Note that the sub-segment pairs (σ, τ) in bold are those confirming the translation hypothesis T, while the rest contradict some parts of the hypothesis. For the word *Machine* (which corresponds to word position 5), there is only one sub-segment pair confirming it *("Automàtica", "Automatic")* with length 1, and no one with lengths 2 and 3. Therefore, we have that:

$$\mathrm{conf}_1(5, S, T, M) = \{(\text{“Automàtica”}, \text{“Automatic”})\}$$

$$\mathrm{conf}_2(5, S, T, M) = \emptyset$$

$$\mathrm{conf}_3(5, S, T, M) = \emptyset$$

In addition, we have that the sub-segments τ in $\mathrm{seg}_*(T)$ covering the word *Automatic* for lengths in $[1, 3]$ are:

$$\{\tau : \tau \in \mathrm{seg}_1(T) \ \wedge \ j \in \mathrm{span}(\tau, T)\} = \{\text{“Automatic”}\}$$

$$\{\tau : \tau \in \mathrm{seg}_2(T) \ \wedge \ j \in \mathrm{span}(\tau, T)\} = \{\text{“the Automatic”}, \text{“Automatic Translation”}\}$$

$$\{\tau : \tau \in \mathrm{seg}_3(T) \ \wedge \ j \in \mathrm{span}(\tau, T)\} = \{\text{“for the Automatic”}, \text{“the Automatic Translation”}\}$$

Therefore, the resulting positive features for this word would be:

$$\mathrm{Pos}_1(5, S, T, M) = \frac{\mathrm{conf}_3(5, S, T, M)}{\{\tau : \tau \in \mathrm{seg}_1(T) \ \wedge \ j \in \mathrm{span}(\tau, T)\}} = \frac{1}{1}$$

$$\mathrm{Pos}_2(5, S, T, M) = \frac{\mathrm{conf}_2(5, S, T, M)}{\{\tau : \tau \in \mathrm{seg}_2(T) \ \wedge \ j \in \mathrm{span}(\tau, T)\}} = \frac{0}{2}$$

$$\mathrm{Pos}_3(5, S, T, M) = \frac{\mathrm{conf}_3(5, S, T, M)}{\{\tau : \tau \in \mathrm{seg}_3(T) \ \wedge \ j \in \mathrm{span}(\tau, T)\}} = \frac{0}{2}$$

A second collection of features, which use the information about the translation frequency between the pairs of sub-segments in M is also used. This information is not available for MT, but it is for the bilingual concordancer we have used (see Section 3). This frequency determines how often σ is translated as τ and, therefore, how reliable this translation is. We define $\mathrm{Pos}_n^{\mathrm{freq}}$ to obtain these features as:

$$\mathrm{Pos}_n^{\mathrm{freq}}(j, S, T, M) = \sum_{\forall(\sigma,\tau)\in \mathrm{conf}_n(j,S,T,M)} \frac{\mathrm{occ}(\sigma, \tau, M)}{\sum_{\forall(\sigma,\tau')\in M} \mathrm{occ}(\sigma, \tau', M)}$$

where function $\mathrm{occ}(\sigma, \tau, M)$ returns the number of occurrences in M of the sub-segment pair (σ, τ).

Following the running example, we may have an alternative and richer source of bilingual information, such as a sub-segmental translation memory, which contains 99 occurrences of word *Automàtica* translated as *Automatic*, as well as the following alternative translations: *Machine* (11 times), and *Mechanic* (10 times). Therefore, the positive feature using these frequencies for sub-segments of length 1 would be:

$$\mathrm{Pos}_1^{\mathrm{freq}}(5, S, T, M) = \frac{99}{99 + 11 + 10} = 0.825$$

Both positive features, $\mathrm{Pos}(\cdot)$ and $\mathrm{Pos}^{\mathrm{freq}}(\cdot)$, are computed for t_j for all the values of sub-segment length $n \in [1, L]$. In addition, they can be computed for both $M_{S\rightarrow T}$ and $M_{T\rightarrow S}$; this yields $4L$ positive features in total for each word t_j.

2.2 Negative features

The negative features, i.e. those features that help to identify words that should be post-edited in the translation hypothesis T, are also based on sub-segment translations $(\sigma, \tau) \in M$, but they are used in a different way. Negative features use those sub-segments τ that fit two criteria: (a) they are the translation of a sub-segment σ from S but are not sub-segments of T; and (b) when they are aligned to T using the edit-distance algorithm (Wagner and Fischer, 1974), both their first word θ_1 and last

word $\theta_{|\tau|}$ can be aligned, therefore delimiting a sub-segment τ' of T. Our hypothesis is that those words t_j in τ' which cannot be aligned to τ are likely to need postediting. We define our negative feature collection $\mathrm{Neg}_{mn'}$ as:

$$\mathrm{Neg}_{mn'}(j,S,T,M) = \sum_{\forall \tau \in \mathrm{NegEvidence}_{mn'}(j,S,T,M)} \frac{1}{\mathrm{alignmentsize}(\tau,T)}$$

where $\mathrm{alignmentsize}(\tau,T)$ returns the length of the sub-segment τ' delimited by τ in T. Function $\mathrm{NegEvidence}_{mn'}(\cdot)$ returns the set of sub-segments τ of T that are considered negative evidence and is defined as:

$$\mathrm{NegEvidence}_{mn'}(j,S,T,M) = \{\tau : (\sigma,\tau) \in M \wedge \sigma \in \mathrm{seg}_m(S) \wedge |\tau'| = n' \wedge \tau \notin \mathrm{seg}_*(T) \wedge \mathrm{IsNeg}(j,\tau,T)\}$$

In this function length constraints are set so that sub-segments σ take lengths $m \in [1,L]$. While for the positive features, only the length of τ was constrained, the experiments carried out by Esplà-Gomis et al. (2015b) indicate that for the negative features, it is better to constrain also the length of σ. On the other hand, the case of the sub-segments τ is slightly different: n' does not stand for the length of the sub-segments, but the number of words in τ which are aligned to T.[7] Function $\mathrm{IsNeg}(\cdot)$ defines the set of conditions required to consider a sub-segment τ a negative evidence for word t_j:

$$\mathrm{IsNeg}(j,\tau,T) = \exists j', j'' \in [1,|T|] : j' < j < j'' \wedge \mathrm{aligned}(t_{j'},\theta_1) \wedge \mathrm{aligned}(t_{j''},\theta_{|\tau|}) \wedge \nexists \theta_k \in \mathrm{seg}_1(\tau) : \mathrm{aligned}(t_j,\theta_k)$$

where $\mathrm{aligned}(X,Y)$ is a binary function that checks whether words X and Y are aligned or not.

For our running example, only two sub-segment pairs (σ,τ) fit the conditions set by function $\mathrm{IsNeg}(j,\tau,T)$ for the word *Automatic*: *("la Traducció", "the Translation")*, and *("la Traducció Automàtica", "the Machine Translation")*. As can be seen, for both (σ,τ) pairs, the words *the* and *Translation* in the sub-segments τ can be aligned to the words in positions 4 and 6 in T, respectively, which makes the number of words aligned $n' = 2$. In this way, we would have the evidences:

$$\mathrm{NegEvidence}_{2,2}(5,S,T,M) = \{\text{"the Translation"}\}$$

$$\mathrm{NegEvidence}_{3,2}(5,S,T,M) = \{\text{"the Machine Translation"}\}$$

As can be seen, in the case of sub-segment $\tau =$ "*the Translation*", these alignments suggest that word *Automatic* should be removed, while for the sub-segment $\tau =$ *the Machine Translation*" they suggest that word *Automatic* should be replaced by word *Machine*. The resulting negative features are:

$$\mathrm{Neg}_{2,2}(5,S,T,M) = \tfrac{1}{3}$$

$$\mathrm{Neg}_{3,2}(5,S,T,M) = \tfrac{1}{3}$$

Negative features $\mathrm{Neg}_{mn'}(\cdot)$ are computed for t_j for all the values of SL sub-segment lengths $m \in [1,L]$ and the number of TL words $n' \in [2,L]$ which are aligned to words θ_k in sub-segment τ. Note that the number of aligned words between T and τ cannot be smaller than 2 given the constraints set by function $\mathrm{IsNeg}(j,\tau,T)$. This results in a collection of $L \times (L-1)$ negative features. Obviously, for these features only $M_{S\to T}$ is used, since in $M_{T\to S}$ all the sub-segments τ can be found in T.

3 Experiments

This section describes the dataset provided for the word-level MTQE sub-task and the results obtained by our method on these datasest. This year, the task consisted in measuring the word-level MTQE on a collection of segments in Spanish that had been obtained through machine translation from English. The organisers provided a dataset consisting of:

- *training set*: a collection of 11,272 segments in English (S) and their corresponding machine translations in Spanish (T); for every word in T, a label was provided: BAD for the words to be post-edited, and GOOD for those to be kept unedited;
- *development set*: 1,000 pairs of segments (S,T) with the corresponding MTQE labels that can be used to optimise the binary classifier trained by using the training set;
- *test set*: 1,817 pairs of segments (S,T) for which the MTQE labels have to be estimated with the binary classifier trained on the training and the development sets.

[7] That is, the length of longest common sub-segment of τ and T.

3.1 Binary classifier

A *multilayer perceptron* (Duda et al., 2000, Section 6) was used for classification, as implemented in Weka 3.6 (Hall et al., 2009), following the approach by Esplà-Gomis et al. (2015b). A subset of 10% of the training examples was extracted from the training set before starting the training process and used as a validation set. The weights were iteratively updated on the basis of the error computed in the other 90%, but the decision to stop the training (usually referred as the convergence condition) was based on this validation set, in order to minimise the risk of overfitting. The error function used was based on the the optimisation of the metric used for ranking, i.e. the F_1^{BAD} metric.

Hyperparameter optimisation was carried out on the development set, by using a grid search (Bergstra et al., 2011) in order to choose the hyperparameters optimising the results for the metric to be used for comparison, F_1 for class *BAD*:

- *Number of nodes in the hidden layer*: Weka (Hall et al., 2009) makes it possible to choose from among a collection of predefined network designs; the design performing best in most cases happened to have a single hidden layer containing the same number of nodes in the hidden layer as the number of features.
- *Learning rate*: this parameter allows the dimension of the weight updates to be regulated by applying a factor to the error function after each iteration; the value that best performed for most of our training data sets was 0.1.
- *Momentum*: when updating the weights at the end of a training iteration, momentum smooths the training process for faster convergence by making it dependent on the previous weight value; in the case of our experiments, it was set to 0.03.

3.2 Evaluation

As already mentioned, two configurations of our system were submitted: one using only the features defined in Section 2, and one combining them with the baseline features. In order to obtain our features we used two sources of bilingual information, as already mentioned: MT and a bilingual concordancer. As explained above, for our experiments we used two MT systems which are freely available on the Internet: Apertium and Google Translate. The bilingual concordancer *Reverso Context* was also used for translating sub-segments. Actually, only the sub-sentential translation memory of this system was used, which provides the collection of TL translation alternatives for a given SL subsegment, together with the number of occurrences of the sub-segments pair in the translation memory.

Four evaluation metrics were proposed for this task:

- The precision P^c, i.e. the fraction of instances correctly labelled among all the instances labelled as c, where c is the class assigned (either GOOD or BAD in our case);
- The recall R^c, i.e. the fraction of instances correctly labelled as c among all the instances that should be labelled as c in the test set;
- The F_1^c score, which is defined as

$$F_1^c = \frac{2 \times P^c \times R^c}{P^c + R^c};$$

  although the F_1^c score is computed both for GOOD and for BAD, it is worth noting that the F_1 score for the less frequent class in the data set (label BAD, in this case) is used as the main comparison metric;
- The F_1^w score, which is the version of F_1^c weighted by the proportion of instances of a given class c in the data set:

$$F_1^w = \frac{N^{\mathrm{BAD}}}{N^{\mathrm{TOTAL}}} F_1^{\mathrm{BAD}} + \frac{N^{\mathrm{GOOD}}}{N^{\mathrm{TOTAL}}} F_1^{\mathrm{GOOD}}$$

  where N^{BAD} is the number of instances of the class BAD, N^{GOOD} is the number of instances of the class GOOD, and N^{TOTAL} is the total number of instances in the test set.

3.3 Results

Table 1 shows the results obtained by our system, both on the development set during the training phase and on the test set. The table also includes the results for the baseline system as published by the organisers of the shared task, which uses the baseline features provided by them and a standard logistic regression binary classifier.

As can be seen in Table 1, the results obtained on the development set and the test set are quite similar and coherent, which highlights the robustness of the approach. The results obtained clearly outperform the baseline on the main evaluation metric (F_1^{BAD}). It is worth noting that, on this metric, the

Data set	System	P^{BAD}	R^{BAD}	F_1^{BAD}	P^{GOOD}	R^{GOOD}	F_1^{GOOD}	F_1^w
development set	SBI	31.2%	63.7%	41.9%	88.5%	66.7%	76.1%	69.5%
	SBI+baseline	33.4%	60.9%	43.1%	88.5%	71.1%	78.8%	72.0%
test set	baseline	—	—	16.8%	—	—	88.9%	75.3%
	SBI	30.8%	63.9%	41.5%	88.8%	66.5%	76.1%	69.5%
	SBI+baseline	32.6%	63.6%	43.1%	89.1%	69.5%	78.1%	71.5%

Table 1: Results of the two systems submitted to the WMT 2015 sub-task on word-level MTQE: the one using only sources of bilingual information (SBI) and the one combining these sources of information with the baseline features (SBI+baseline). The table also includes the results of the baseline system proposed by the organisation; in this case only the F_1 scores are provided because, at the time of writing this paper, the rest of metrics remain unpublished.

SBI and SBI+baseline submissions scored first and third among the 16 submissions to the shared task.[8] The submission scoring second obtained very similar results; for F_1^{BAD} it obtained 43.05%, while our submission obtained 43.12%. On the other hand, using the metric F_1^w for comparison, our submissions ranked 10 and 11 in the shared task, although it is worth noting that our system was optimised using only the F_1^{BAD} metric, which is the one chosen by the organisers for ranking submissions.

4 Concluding remarks

In this paper we described the submissions of the *UAlacant* team for the sub-task 2 in the MTQE shared task of the WMT 2015 (word-level MTQE). Our submissions, which were ranked first and third, used online available sources bilingual of information in order to extract relations between the words in the original SL segments and their TL machine translations. The approach employed is aimed at being system-independent, since it only uses resources produced by external systems. In addition, adding new sources of information is straightforward, which leaves considerable room for improvement. In general, the results obtained support the conclusions obtained by Esplà-Gomis et al. (2015b) regarding the feasibility of this approach and its performance.

Acknowledgements

Work partially funded by the Spanish Ministerio de Ciencia e Innovación through project TIN2012-32615 and by the European Commission through project PIAP-GA-2012-324414 (AbuMaTran). We specially thank Reverso-Softissimo and Prompsit Language Engineering for providing the access to the Reverso Context concordancer, and to the University Research Program for Google Translate that granted us access to the Google Translate service.

[8] http://www.quest.dcs.shef.ac.uk/wmt15_files/results/task2.pdf

References

J.S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. 2011. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04.

R. O. Duda, P. E. Hart, and D. G. Stork. 2000. *Pattern Classification*. John Wiley and Sons Inc., 2nd edition.

M. Esplà-Gomis, F. Sánchez-Martínez, and M.L. Forcada. 2012. UAlacant: Using online machine translation for cross-lingual textual entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval 2012), pages 472–476, Montreal, Canada.

M. Esplà-Gomis, F. Sánchez-Martínez, and M. L. Forcada. 2015a. Target-language edit hints in CAT tools based on TM by means of MT. *Journal of Artificial Intelligence Research*, 53:169–222.

M. Esplà-Gomis, F. Sánchez-Martínez, and M. L. Forcada. 2015b. Using on-line available sources of bilingual information for word-level machine translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Associtation for Machine Translation*, pages 19–26, Antalya, Turkey.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, 11(1):10–18.

L. Specia and R. Soricut. 2013. Quality estimation for machine translation: preface. *Machine Translation*, 27(3-4):167–170.

L. Specia, D. Raj, and M. Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

R.A. Wagner and M.J. Fischer. 1974. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173.