10.12871/CLICIT2014128

Online and Multitask Learning for Machine Translation Quality Estimation in Real-world Scenarios

José G. C. de Souza[(1,2)] **Marco Turchi**[(1)]
Antonios Anastasopoulos[(3)] **Matteo Negri**[(1)]
[(1)] FBK - Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
[(2)] University of Trento, Italy
[(3)] University of Notre Dame, Indiana, USA
{desouza,turchi,negri}@fbk.eu
aanastas@nd.edu

Abstract

English. We investigate the application of different supervised learning approaches to machine translation quality estimation in realistic conditions where training data are not available or are heterogeneous with respect to the test data. Our experiments are carried out with two techniques: *online* and *multitask* learning. The former is capable to learn and self-adapt to user feedback, and is suitable for situations in which training data is not available. The latter is capable to learn from data coming from multiple domains, which might considerably differ from the actual testing domain. Two focused experiments in such challenging conditions indicate the good potential of the two approaches.

Italiano. *Questo articolo descrive l'utilizzo di tecniche di apprendimento supervisionato per stimare la qualità della traduzione automatica in condizioni in cui i dati per l'addestramento non sono disponibili o sono disomogenei rispetto a quelli usati per la valutazione. A tal fine si confrontano due approcci: online e multitask learning. Il primo consente di apprendere da feedback degli utenti, dimostrandosi adatto a situazioni di assenza di dati. Il secondo consente l'apprendimento da dati provenienti da più dominii, anche molto diversi da quello in cui il sistema verrà valutato. I risultati di due esperimenti in tali scenari suggeriscono l'efficacia di entrambi gli approcci.*

1 Introduction

Quality Estimation (QE) for Machine Translation (MT) is the task of estimating the quality of a translated sentence at run-time and without access to reference translations (Specia et al., 2009; Soricut and Echihabi, 2010; Bach et al., 2011; Specia, 2011; Mehdad et al., 2012; C. de Souza et al., 2013; C. de Souza et al., 2014a). As a quality indicator, in a typical QE setting, automatic systems have to predict either the time or the number of editing operations (*e.g.* in terms of HTER[1]) required to a human to transform the translation into a syntactically/semantically correct sentence. In recent years, QE gained increasing interest in the MT community as a possible way to: *i)* decide whether a given translation is good enough for publishing as is, *ii)* inform readers of the target language only whether or not they can rely on a translation, *iii)* filter out sentences that are not good enough for post-editing by professional translators, or *iv)* select the best translation among options from multiple MT and/or translation memory systems.

So far, despite its many possible applications, QE research has been mainly conducted in controlled lab testing scenarios that disregard some of the possible challenges posed by real working conditions. Indeed, the large body of research resulting from three editions of the shared QE task organized within the yearly Workshop on Machine Translation (WMT – (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014)) has relied on simplistic assumptions that do not always hold in real life. These assumptions include the idea that the data available to train QE models is: *i) large* (WMT systems are usually trained over datasets of 800/1000 instances) and *ii) representative* (WMT training and test sets are always drawn from the same domain and are uniformly distributed).

[1]The HTER (Snover et al., 2006) measures the minimum edit distance between the MT output and its manually post-edited version in the [0,1] interval. Edit distance is calculated as the number of edits (word insertions, deletions, substitutions, and shifts) divided by the number of words in the reference. Lower HTER values indicate better translations.

In order to investigate the difficulties of training a QE model in realistic scenarios where such conditions might not hold, in this paper we approach the task in situations where: *i)* training data is not available at all (§2), and *ii)* training instances come from different domains (§3). In these two situations, particularly challenging from the machine learning perspective, we investigate the potential of online and multitask learning methods (the former for dealing with the lack of data, and the latter to cope with data heterogeneity), comparing them with the batch methods currently used.

2 How to obtain a QE model *without* training data?

Our first experiment addresses the problem of building a QE model from scratch, when training data is not available (*i.e.* by only learning from the test set). In this scenario, we apply the online learning protocol as a way to build our model and stepwise refine its predictions by exploiting user feedback on the processed test instances.

In the online framework, differently from the batch mode where the model is built from an available training set, the learning algorithm sequentially processes an unknown sequence of instances $X = x_1, x_2, ..., x_n$, returning a prediction $p(x_i)$ as output at each step. Differences between $p(x_i)$ and the true label $\hat{p}(x_i)$ obtained as feedback are used by the learner to refine the next prediction $p(x_{i+1})$. In our experiment we aim to predict the quality of the suggested translations in terms of HTER. In this scenario:

- The set of instances X is represented by (*source, target*) pairs;
- The prediction $p(x_i)$ is the automatically estimated HTER score;
- The true label $\hat{p}(x_i)$ is the actual HTER score calculated over the target and its post-edition.

At each step of the process, the goal of the learner is to exploit user post-editions to reduce the difference between the predicted HTER values and the true labels for the following (*source, target*) pairs. Similar to (Turchi et al., 2014), we do it as follows:

1. At step *i*, an unlabelled (*source, target*) pair x_i is sent to a feature extraction component. To this aim, we used an adapted version (Shah et al., 2014) of the open-source QuEst tool (Specia et al., 2013). The tool, which implements a large number of features proposed by participants in the WMT QE shared tasks, has been modified to process one sentence at a time;
2. The extracted features are sent to an online regressor, which returns a QE prediction score $p(x_i)$ in the [0,1] interval (set to 0 at the first round of the iteration);
3. Based on the post-edition done by the user, the true HTER label $\hat{p}(x_i)$ is calculated by means of the TERCpp[2] open source tool;
4. The true label is sent back to the online algorithm for a stepwise model improvement. The updated model is then ready to process the following instance x_{i+1}.

Online vs batch algorithms. We compare the results achieved by OnlineSVR (Parrella, 2007)[3] with those obtained by a batch strategy based on the Scikit-learn implementation of Support Vector Regression (SVR).[4] Our goal is to check to what extent the online approach (which learns from scratch from the test set) can approximate the batch results obtained, in more favourable conditions, with different amounts of training data.

Feature set. Our feature set consists of the seventeen features proposed in (Specia et al., 2009). These features, fully described in (Callison-Burch et al., 2012), take into account the complexity of the source sentence (*e.g.* number of tokens, number of translations per source word) and the fluency of the translation (*e.g.* language model probabilities). The results of previous WMT QE shared tasks have shown that these baseline features are particularly competitive in the regression task.

Performance indicator. Performance is measured by computing the Mean Absolute Error (MAE), a metric for regression problems also used in the WMT QE shared tasks. The MAE is the average of the absolute errors $e_i = |f_i - y_i|$, where f_i is the prediction of the model and y_i is the true value for the i^{th} instance.

Dataset. Our dataset is drawn from the WMT12 English-Spanish corpus and consists of: *i)* three training sets of different size (200, 600, and 1500

[2] goo.gl/nkh2rE

[3] http://www2.imperial.ac.uk/~gmontana/onlinesvr.htm

[4] http://scikit-learn.org/

instances) used to build the batch models, and *ii)* one "test" set of 754 instances used to build the online models and compare the results obtained with the two strategies. The HTER labels used as feedback by the online approach are calculated using the post-edited version of the target sentences, which is also provided in the WMT12 dataset.

Results. Table 1 reports the MAE results achieved by SVR models (batch strategy) obtained from the three training sets, and the result achieved by the OnlineSVR model (online strategy) obtained by learning only from the test set (since the model is always trained on the same test set, this result of 13.5% MAE is always the same).

As can be seen from the table, similar MAE values show a similar behaviour for the two strategies. This holds even when the batch method can take advantage of the largest dataset to learn from (1500 instances, twice the size of the data used by OnlineSVR).[5] For the batch method, this is an ideal condition not only due to the large amount of data to learn from, but also due to the high homogeneity of the training and test sets (indeed, WMT data come from the same domain and are uniformly distributed). In spite of this, when moving from 600 to 1500 training instances, SVR performance gets stable to a value (12.5% MAE) that is not significantly better than the performance achieved by OnlineSVR. Finally, it's worth noting that, since they are calculated over the entire test set, OnlineSVR results can be highly affected by completely wrong predictions returned for the first instances (recall that at the first step the model returns 0 as a default value). These results, particularly interesting from an application-oriented perspective, indicate the potential of online learning to deal with situations in which training data is not available.

Train	Test	SVR	OnlineSVR
200	754	13.2	13.5*
600	754	12.7	13.5*
1500	754	12.7	13.5*

Table 1: QE performance (MAE) of three batch models (SVR) built from different amounts of training data, and one online model (OnlineSVR) that only learns from the test set.

[5]The online results marked with the "*" symbol are NOT statistically significant compared to the corresponding batch model. Statistical significance at $p \leq 0.005$ has been calculated with approximate randomization (Yeh, 2000).

3 How to obtain a QE model from *heterogeneous* training/test data?

The dominant QE framework presents some characteristics that can limit models' applicability in real-world scenarios. First, the scores used as training labels (HTER, time) are costly to obtain because they are derived from manual post-editions of MT output. Such requirement makes it difficult to develop models for domains in which there is a limited amount of labelled data. Second, the learning methods currently used assume that training and test data are sampled from the same distribution. Though reasonable as a first evaluation setting to promote research in the field, this controlled scenario is not realistic because different data in real-world applications might be post-edited by different translators whose different attitudes have to be modelled (Cohn and Specia, 2013; Turchi et al., 2013; Turchi et al., 2014), the translations might be generated by different MT systems and the documents being translated might belong to different domains or genres.

To overcome these limitations, which represent a major problem for current batch approaches, a reasonable research objective is to exploit techniques that: *i)* allow domains and distributions of features to be different between training and test data, and *ii)* cope with the scarce amount of training labels by sharing information across domains.

In our second experiment we investigate the use of techniques that can exploit training instances from different domains to learn a QE model for a specific target domain for which there is a small amount of labelled data. As suggested in (C. de Souza et al., 2014b) this problem can be approached as a *transfer learning* problem in which the knowledge extracted from one or more source tasks is applied to a target task (Pan and Yang, 2010). *Multitask learning*, a special case of transfer learning, uses domain-specific training signals of related tasks to improve model generalization (Caruana, 1997). Although it was not originally thought for transferring knowledge to a new task, MTL can be used to achieve this objective due to its capability to capture task relatedness, which is important knowledge that can be applied to a new task (Jiang, 2009). When applied to domain adaptation, the approach is transformed in a standard learning problem by augmenting the source and target feature set. The feature space is transformed to be a cross-product of the features of

the source and target domains augmented with the original target domain features. In *supervised* domain adaptation, out-of-domain labels and a small amount of available in-domain labelled data are exploited to train a model (Daumé III, 2007). This is different from the *semi-supervised* case, in which in-domain labels are not available.

Multitask vs single task algorithms. Our approach falls in the supervised domain adaptation framework, for which we apply the Robust MTL approach (RMTL – (Chen et al., 2011)). Our goal is to check to what extent this approach can improve over single task learning strategies. To this aim, RMTL is compared with: *i)* a regressor built only on the available in-domain data (SVR In-domain), and *ii)* a regressor trained by pooling together the training data from all domains, without any kind of task relationship notion (SVR Pooling). These two regressors are built using the implementation of Scikit-learn (Pedregosa et al., 2011).

Feature set and performance indicator. In this experiment we use the same feature set (Specia et al., 2009) and the same performance indicator (MAE) used in §2.

Dataset. Our experiments focus on the English-French language pair and encompass three very different domains: newswire text (henceforth News), transcriptions of Technology Entertainment Design talks (TED) and Information Technology manuals (IT). Such domains represent a challenging combination for adaptive systems since they come from very different sources spanning speech and written discourse (TED and News/IT, respectively) as well as a very well defined and controlled vocabulary in the case of IT. Each domain is composed of 363 tuples formed by the source sentence in English, the French translation produced by an MT system and a human post-edition of the translated sentence. For each pair (translation, post-edition) we compute the HTER to be used as label. For the three domains we use half of the data for training (181 instances) and the other half for testing (182 instances). The reduced amount of instances for training contrasts with the 800 or more instances of the WMT evaluation campaigns and is closer to real-world applications where the availability of large training sets is far from being guaranteed. The sentence tuples for the first two domains were randomly sampled from the Trace corpus[6]. The translations were generated by two different MT systems and post-edited by up to four different translators as described in (Wisniewski et al., 2013). The IT texts come from a software user manual translated by a statistical MT system based on the state-of-the-art phrase-based Moses toolkit (Koehn et al., 2007) trained on about 2M parallel sentences. The post-editions were collected from one professional translator operating in real working conditions with the MateCat tool (Federico et al., 2014).

Results. Table 2 reports the MAE results achieved by the three models (RMTL, SVR In-domain, SVR Pooling). As can be seen from the table, RMTL always outperforms the other methods with statistically significant improvements. These results provide a strong evidence about the higher suitability of multitask learning to deal with real-world contexts that require robust methods to cope with scarce and heterogeneous training data.

Method	TED	News	IT
30 % of training data (54 instances)			
SVR In-Domain	0.2013	0.1753	0.2235
SVR Pooling	0.1962	0.1899	0.2201
RMTL	**0.1946**	**0.1685**	**0.2162**
50% of training data (90 instances)			
SVR In-Domain	0.1976	0.1711	0.2183
SVR Pooling	0.1951	0.1865	0.2191
RMTL	**0.1878**	**0.1653**	**0.2119**
100% of training data (181 instances)			
SVR In-Domain	0.1928	0.1690	0.2081
SVR Pooling	0.1927	0.1849	0.2203
RMTL	**0.1846**	**0.1653**	**0.2075**

Table 2: Average performance (MAE) of fifty runs of the models (multitask RMTL and the single-task SVR In-domain and SVR Pooling) on 30, 50 and 100 percent of training data.

4 Conclusion

We investigated the problem of training reliable QE models in particularly challenging conditions from the learning perspective. Two focused experiments have been carried out by applying: *i)* online learning to cope with the lack of training data, and *ii)* multitask learning to cope with heterogeneous training data. The positive results of our experiments suggest that the two paradigms should be further explored (and possibly combined) to overcome the limitations of current methods and make QE applicable in real-world scenarios.

[6] http://anrtrace.limsi.fr/trace_postedit.tar.bz2

Acknowledgments

This work has been partially supported by the EC-funded project MateCat (ICT-2011.4.2-287688).

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: a Method for Measuring Machine Translation Confidence. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 211–219. The Association for Computer Linguistics.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, WMT-2013, pages 1–44, Sofia, Bulgaria.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

José G. C. de Souza, Christian Buck, Marco Turchi, and Matteo Negri. 2013. FBK-UEdin Participation to the WMT13 Quality Estimation Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 352–358, Sofia, Bulgaria, August. Association for Computational Linguistics.

José G. C. de Souza, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014a. FBK-UPV-UEdin Participation in the WMT14 Quality Estimation Shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

José G. C. de Souza, Marco Turchi, and Matteo Negri. 2014b. Machine Translation Quality Estimation Across Domains. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 409–420, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT'12)*, pages 10–51, Montréal, Canada.

Rich Caruana. 1997. Multitask Learning. *Machine learning*, 28(1):41–75.

Jianhui Chen, Jiayu Zhou, and Jieping Ye. 2011. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, page 42, New York, New York, USA. ACM Press.

Trevor Cohn and Lucia Specia. 2013. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL-2013, pages 32–42, Sofia, Bulgaria.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Conference of the Association for Computational Linguistics (ACL)*.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. THE MATECAT TOOL. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Jing Jiang. 2009. Multi-Task Transfer Learning for Weakly-Supervised Relation Extraction. In *ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, number August, pages 1012–1020.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 171–180, Montréal, Canada.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October.

Francesco Parrella. 2007. Online support vector regression. *Master's Thesis, Department of Information Science, University of Genoa, Italy*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kashif Shah, Marco Turchi, and Lucia Specia. 2014. An Efficient and User-friendly Tool for Machine Translation Quality Estimation. In *Proceedings of the 9^th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proceedings of the 48^{th} Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 612–621.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13^{th} Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 28–35, Barcelona, Spain.

Lucia Specia, Kashif Shah, José G.C. de Souza, and Trevor Cohn. 2013. QuEst - A Translation Quality Estimation Framework. In *Proceedings of the 51^{st} Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL-2013, pages 79–84, Sofia, Bulgaria.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. pages 73–80.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proceedings of the 8^{th} Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria.

Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Baltimore, Maryland, June. Association for Computational Linguistics.

Guillaume Wisniewski, Anil Kumar Singh, Natalia Segal, and François Yvon. 2013. Design and Analysis of a Large Corpus of Post-Edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-Edition. In *Machine Translation Summit XIV*, pages 117–124.

Alexander Yeh. 2000. More Accurate Tests for the Statistical Significance of Result Differences. In *Proceedings of the 18th conference on Computational linguistics (COLING 2000) - Volume 2*, pages 947–953, Saarbrucken, Germany.