Large Scale Translation Quality Estimation

Miguel Rios, Serge Sharoff
Centre for Translation Studies
University of Leeds
m.riosgaona,s.sharoff@leeds.ac.uk

Abstract

This study explores methods for developing a large scale Quality Estimation framework for Machine Translation. We expand existing resources for Quality Estimation across related languages by using different transfer learning methods. The transfer learning methods are: Transductive SVM, Label Propagation and Self-taught Learning. We use transfer learning methods on the available labelled datasets, e.g. en-es, to produce a range of Quality Estimation models for Romance languages, while also adapting for subtitling as a new domain. The Self-taught Learning method shows the most promising results among the used techniques.

1 Introduction

A common problem with automatic metrics for Machine Translation (MT) evaluation, such as BLEU (Papineni et al., 2002), is the need to have reference human translations. Also such metrics work best on a corpus of sentences, while they are not informative for evaluation of individual sentences (Specia et al., 2009). The aim of Quality Estimation (QE) is to predict a quality score for sentences output by MT without reference translations, for example, to judge whether they provide a suitable basis for Post-Editing by the human translator or it is better to ask the human to translate this sentence from scratch. The QE task can be framed as a classification or a regression problem, where most of the methods for QE rely on supervised Machine Learning (ML) algorithms.

The WMT evaluation campaigns (Bojar et al., 2014) goal is to create a framework to test the performance of participating systems for the QE task. The WMT organizers provide the datasets for training and testing new proposed automatic QE approaches. However, the existing training data is only available for a limited number of languages. For example, in the WTM 2014 the available pairs were en-es and en-de (throughout the paper we will be using the two-letter ISO codes to indicate the languages). Most of the final MT users and projects need a wider variety of source and target languages for evaluation.

Turchi and Negri (2014) propose an automatic approach to produce training data for QE and tackle the problem of scarce training resources. The approach is based on features across the MT output, the post edited version and the human reference translation. The method produces a classifier for binary estimation by exploiting the characteristics of good translations and their relation with the post-editing process. The produced data is labelled with a binary quality score (i.e. god or bad translation) to overcome biases on the annotation.

On the other hand, Birch et al. (2008) propose a large scale study on the performance of 110 European language pairs over Europarl. The study is based on the measuring the contribution of different features between language pairs that improve or are irrelevant to the performance of an MT system. The features consist of complexity indicators of morphology, language relatedness given word similarity, number of reordering between language pairs and number of reorderings over alignments. Overall, closely related languages showed the best potential for SMT. However, this study is mainly based on standard automatic evaluation metrics such as BLEU.

In this ongoing work, we propose a way to produce a number of evaluation pairs for the QE task by utilising relatedness between languages, for example, by producing a QE evaluation for the en-pt pair from an existing en-es training set. More specifically we will study the use of different transfer learning methods to transfer classifiers across related languages. Our intuition is that sentences with similar quality scores are close or share a lower-dimensional space in terms of features across related languages. In other words, good/bad quality sentences might show similar characteristics between the available training data (e.g. en-es) and unknown data (e.g en-pt). This makes possible the training of a classification algorithm to predict QE by sharing information from the available dataset into unknown datasets.

We show preliminary results on transferring training data from en-es European Parliament (Europarl) domain to en-es, en-it and en-pt in the subtitling domain. The transfer learning method that shows promising results is the one based on dimensionality reduction of the input. However, this method is sensible to the distribution of classes of the training dataset, where it tends to predict the majority of the training class. In addition, we provide further directions into transfer training data based on the similarity of related languages for source languages that are not present in the original WMT QE datasets, but also to tackle the unbalanced training dataset. We use a simple heuristic for assigning possible labels for the unlabelled data based on edit distance scores between available reference translations and MT outputs. However, this simple heuristic hurts the performance of the methods, where a more appropriate way of adding similarity information is as an indicator for domain shift.

2 Background

Methods for QE are commonly based on computing similarity scores and information supplied by the MT decoding process between source and machine translations. These sources of information are used as features to train a supervised ML algorithm to predict QE scores. Specia et al. (2013) develop the standard baseline framework for QE based on features that attempt to quantify the complexity of a segment to be translated. Other previous works extend the baseline framework by adding complex features between the source and machine translations. For example, syntax information of tree labels counts (Avramidis, 2014), information to quantify the acts of translation between any two datasets with respect to a reference in the same domain (Bicici and Way, 2014) and word alignment, word posterior probabilities and diversity scores features (Camargo de Souza et al., 2014).

Beck et al. (2014) use multi-task learning techniques to improve QE by sharing information among different domains. However, the QE task is only applied to certain language pairs. On the other hand, de Souza et al. (2015) integrate QE into a CAT tool with online learning to constantly train the quality prediction model. This method can be used to extract QE training data or prediction models for several domains and languages.

Transfer learning aims to transfer information learned in one or more source tasks (e.g. labelled dataset) and use it to improve learning in a related target task (e.g. unlabelled dataset) (Pan and Yang, 2010). In our case the labelled dataset comes from a QE training set for an existing language pair, while unlabelled datasets are either for the same pair, but in a completely different domain, or for another language pair.

2.1 Transductive Support Vector Machine

Transductive Support Vector Machine (TSVM) takes into consideration a particular test dataset and tries to minimise errors only on those particular instances (Vapnik, 1995). The particular test dataset is added into the training dataset without labels. The TSVM learns a large margin hyperplane classifier using labelled training data, but at the same time it forces that hyperplane to be far from the unlabelled data. The TSVM considers f that maps inputs x to outputs y. However, TSVM does not construct a function f where the output of the transduction algorithm is a vector of labels, and the method transfers the information from labelled instances to the unlabelled.

2.2 Label Propagation

Label propagation (Zhu and Ghahramani, 2002) is based on a graph that connects similar instances. The nodes labels (i.e. instances) propagate to neighbouring nodes given proximity. This model resembles the k-NN nearest neighbours where closer data points tend to have similar labels. The l labelled training examples $\{(x_1, y_1), (x_2, y_2)..., (x_l, y_l)\}$ and the u unlabelled training examples $\{x_1, x_2, ..., x_u\}$, where the Y classes are known. Label propagation estimates the Y_u given the training examples. The method creates a fully connected graph where the nodes are all the labelled and unlabelled instances. The edges are weighted based on the euclidean distance between the nodes where the closer nodes have a larger weight value. The nodes have soft labels that are propagated thorough all the edges modifying the unlabelled instances, and the larger the weight the easier is to propagate the label across the graph.

2.3 Self-taught Learning

Raina et al. (2007) propose a semi-supervised transfer learning method based on using labelled and unlabelled data. However, this method does not assume that the unlabelled dataset is drawn from the same distribution as the labelled. The unlabelled data is used to learn a lower dimensional feature representation of the inputs. With this representation new instances can be classified in the lower dimensional space. The unlabelled data is used for dimensionality reduction of the labelled dataset, which is commonly used with sparse high dimensional data.

The transfer learning problem algorithm is defined as:

- l training examples $\{(x_1, y_1), (x_2, y_2)..., (x_l, y_l)\}, y \in \mathcal{Y}$;
  where Y is the output.
- u unlabelled examples $\{x_1, x_2, ..., x_u\}$.
- Learn the higher-level representation by dimensionality reduction by using sparse-coding.
- Compute new labelled training dataset with new representation $\hat{x_l}$.
- Use standard classification methods with new training dataset.

3 Methodology

In this section, we describe the QE features and the transfer learning setup. We use the standard QE baseline features and available implementations of transfer learning methods making the experiments easy to reproduce. The QE task (Bojar et al., 2014) considers word-level, sentence-level and document-level estimation. The types of annotation (i.e. labels) for the predicted output scores and ranks consist in:

Post-editing effort The perceived effort of a translator to edit a sentence scored with quality labels such as:
1 = perfect translation, no post-editing needed at all;
2 = near miss translation: translation contains maximum of 2-3 errors, and possibly additional errors that can be easily fixed (capitalisation, punctuation);
3 = very low quality translation, cannot be easily fixed.

HTER The minimum edit distance score between the machine translation and its manually post-edited version in [0,1].

Post-editing time The real valued estimate of the time (in milliseconds) it takes a translator to post-edit the translation.

We focus our experiments on the sentence-level estimation with the labelling based on post-editing effort. However, the transfer learning methods can be applied on every estimation subtask. We chose the 3-way labelling in contrast to the binary classification given that post-editing is a common scenario present in our domain of interest (subtitling). We want to show to the translator sentences with good quality but also sentences that can be saved by a small post-editing effort. In addition, we belied that the 3-way labelling is a straight forward scheme for annotators. Our current experiments only cover the testing of a small number language pairs with manual evaluation.

3.1 QE feature description

The baseline features for QE are defined for the source, target (i.e. MT output) and the translations (i.e. relations between them). The QuEsT framework (Specia et al., 2013) implements different types of features. The features can be divided on different families:

Complexity Indicators Features related to the source text of how complex to translate a sentence can be, such as, number of tokens, language model and average number of translations.

Confidence Indicators Features related to the fluency of the MT output, such as, number of tokens, language model, and number of occurrences of the target word within the target sentence.

Fluency Indicators Features related to the adequacy (meaning preservation) of the translation, such as, ratios of tokens between the source and target, ratio of punctuation and syntactic similarity. The framework also introduces features related to a specific decoding process when available, such as, global score of the system and number of hypotheses in the n-best list.

We use the baseline setup of the framework that consists of 17 baseline features that are language independent.

3.2 Transfer learning setup

We aim to apply transfer learning, when texts in related languages are treated as unlabelled out-of-domain data. For example, the available en-es labelled dataset is used to transfer information into the unlabelled en-pt to predict QE scores on that unknown language pair. The methods used in this study required as input a small amount of labelled instances and large amounts of unlabelled instances for training. We define three models for transfer information from labelled QE data into unlabelled data. The models are as follows:

TSVM Model based on a Transductive Support Vector Machine.

LP Model based on Label Propagation.

STL Model based on Self-taught learning.

We use SVMlin[1] for training the TSVM, given that is optimised to work with a large number of instances. Our TSVM uses an RBF kernel with no hyper-parameter optimisation. Each instance in the unlabelled dataset is added to the training dataset. This improved training data is used to perform testing. For classification, we implement the one-against-one strategy, and the final decision is given by voting.

For the LP model, we use the implementation from the scikit-learn[2] toolkit with the RBF kernel with no hyper-parameter optimization.

We modify the STL MATALB implementation from the Stanford Deep learning course[3]. The STL model first finds the weights b from the unlabelled x_u dataset by training a sparse autoencoder. The b weights come from the optimisation of the cost function on sparse coding, where one of the components are the basis vectors. This is a technique for dimensionality representation of the input. Second, the model produces a modified training dataset by using the unlabelled b weights on an second autoencoder. The modified training dataset is a lower-dimensional representation of the input features (i.e. QE 17 baseline features). We use the softmax regression as classifier with the default parameters and the modified labelled training dataset.

[1] http://vikas.sindhwani.org/svmlin.html

[2] http://scikit-learn.org/dev/index.html

[3] https://github.com/amaas/stanford_dl_ex

A new test dataset can be predicted by using the weights b to represent the data points into the same lower-dimensional space. We normalize the features with the z-score. However, we do not have access to any development datasets for tuning the x_u autoencoder for our unlabelled language pairs. For the parameter selection of the unlabelled autoenconder, as suggested in (Bergstra and Bengio, 2012), we run a random search over a split of the modified training dataset (90% training, 10% validation) in order to find: the size of the hidden dimension, the sparsity parameter, the weight decay parameter and the sparsity penalty. We run the random search parameter optimisation for each unlabelled language pair, thus learning parameters on each unlabelled language pair.

In addition, we define a model based on Logistic Regression without the aid of any transfer learning as the *baseline*. The baseline is trained with an available dataset (e.g. en-es).

4 Experiments

In this section, we describe the data used to train and evaluate our transfer learning models for related languages pairs. We show results on manual evaluation for different language pairs of the en to Romance languages (es, pt and it). We also show cross validation results for the en-de pair.

4.1 Data description

The labelled data x_l for the pair en-es come from the WMT 2014 QE shared task[4], which consist of 3,816 source and target pairs. The en-es WMT data belong to the proceedings of the European Parliament (Europarl) domain. The distribution of instances for each quality label is: *1-949*, *2-2010* and *3-857*. Our objective is to score sentence-level QE for related languages for the *en-target* translation direction, where we vary the target language.

The unlabelled data consist of subtitles from the **Zoo** corpus. Zoo is a proprietary corpus of subtitles produced by professional translators. We split the Zoo corpus into unlabelled training x_u and testing for each one of the pairs: en-es, en-pt and en-it. We also test the pair en-de Europarl given that labelled data is available with 600 sentences for testing, as well as, a correspondent out-of-domain data with 297 sentences. We use the Moses (Koehn et al., 2007) toolkit with a phrase-based baseline to extract the QE features for the x_l, x_u, and testing. The Zoo dataset used for the SMT baseline is: 80K training sentences, 1K sentences for tuning and 2K sentences for testing. We use the Zoo test 2K sentences for testing our proposed methods. We use fast-align[5], KenLM[6] with a 3-gram language model and Moses with the standard feature set. In addition, we run a small QE manual evaluation over a random sample of 100 sentences from the Zoo test dataset (original 2K sentences) for the pairs: en-es, en-pt and en-it. The annotation is performed by one professional translator for Post-editing effort at sentence level with 3-way labelling. The evaluation metric is the absolute classification accuracy for the 3-way labelling between the QE system prediction and the test random sample.

4.2 Results

Table 1 shows the results on the validation dataset for the parameter optimisation of the STL model.

Table 1: Accuracy results for the validation dataset with the STL model.

Pair Model	en-es	en-pt	en-it
STL	0.56	0.55	0.57

We run the random search for learning parameters on the modified training data $\hat{x}_l$ for each unlabelled dataset, where the number of iterations for each random search is 100. The labelled training set is en-es EuroParl and the unlabelled are: es, it and pt (subtitling domain). Each unlabelled dataset consists of

[4] http://www.statmt.org/wmt14/quality-estimation-task.html

[5] https://github.com/clab/fast_align

[6] https://kheafield.com/code/kenlm/

10K sentences from the Zoo training section. It is worth noticing that the learned hidden dimension for each language pair is: en-es 15, en-pt 9, en-it 13, where the original input dimension is 17 features.

Table 2 shows the accuracy results for each transfer learning method on the test samples. The TSVM shows a poor performance in comparison to the other techniques. A possible reason for this result is the lack of parameter optimisation, in specific the parameter for setting the fraction of positive instances for the unlabelled data. Our models are trained with a very unbalanced dataset. The LP results show a similar behaviour, where we manually set a low gamma parameter in order to change the strong bias of predicting all the instances into one class. However, we are able to optimize parameters for the STL given that the model operates over a transformation of the labelled training dataset.

Table 2: Accuracy results for Transfer Learning models on Romance languages pairs.

Pair Model	en-es	en-pt	en-it
TSVM	0.52	0.30	0.30
LP	0.49	0.26	0.31
STL	**0.53**	**0.48**	**0.49**
Baseline	0.50	0.38	0.33

The **STL** model outperforms both the baseline and other transfer techniques. The pair en-es achieves the best results given that is an instance of domain adaptation between the same translation pairs. The performance difference between the STL model and the baseline for en-es is narrow with the same language pair but with different domains (i.e. WMT and subtitling). However, the other pairs achieve lower results in comparison because they have different domains and labelled training language data.

We vary the number of training instances for the en-pt x_u to test the effect over the labelled data. Table 3 shows the 10-fold cross validation and test results on the variation of unlabelled data for the en-pt pair. The number of unlabelled data used for the variation test is as follows: 500, 1K, 10K and 20K. The variation of unlabelled instances marginally affects the cross validation, but over the test dataset the 10K dataset improves the results. However, the balance of labelled instances highly affects the space induced by the autoencoder.

The labelled dataset tends to have the majority of instances into the classes *1,2*, where the STL shows a bias on the prediction for the majority of the classes 1 and 2 from the training dataset. In order to tackle the unbalance labelled data, we use a simple heuristic of selecting the missing *3* class instances, where the Levenshtein distance between the available reference translations and MT outputs is over a certain threshold. The examples are tagged as *3* and added into the labelled training data. For the en-pt pair the number of artificial examples is *161* with a threshold of *0.5*. The accuracy result for the validation is 0.56 and the test accuracy is **0.37**. The validation score shows a marginal improvement, but the heuristic hurts the test accuracy. Yang and Eisenstein (2015) use features to characterise multi-domain shift by a binary vector of which instances share a given domain. In our case the instances can share information by computing similarity between the labelled and unlabelled datasets, as well as, the use of dimensionality reduction.

Table 4 shows the accuracy results on en-de Europarl (1400 instances) as the labelled training and

Table 3: Accuracy results 10-fold cross validation and test dataset en-pt for unlabelled data size variation.

Unlabelled data size (sentences)	10-fold cross validation Training	Test
500	0.52	0.39
1K	0.54	0.37
10K	0.54	**0.48**
20K	0.53	0.41
50K	0.53	0.40

en-de subtitling (10K instances) as the unlabelled dataset for the STL model. We use 10-fold cross validation over the modified training dataset because there is no test data available for en-de subtitling. Over the validation dataset the en-de achieves **0.61**, with a hidden dimension of 15.

Table 4: Results 10-fold cross validation for STL on the en-de pair.

Model \ Pair	en-de
STL	0.47
Baseline	**0.48**

Table 5 shows the results of the STL result for the available WMT Europarl data,WMT out-of-domain data and the baseline.

Table 5: Accuracy results for en-de WMT data.

Model \ Pair	en-de WMT Europarl	en-de WMT out-of-domain
STL	**0.51**	**0.49**
Baseline	0.44	0.41

The distribution of classes on the labelled en-de dataset is: *1-317*, *2-522* and *3-561*. This labelled dataset shows to be balanced in comparison with the en-es. The STL test results outperforms the baseline for the Europarl and out of domain, but the results are lower for the cross validation. The STL model assigns predictions to classes as follows: *1-14.14%*, *2-43.10%* and *3-42.76%*.

5 Future work

We have presented work in progress for developing QE for a large number of language pairs. We use different transfer learning mechanisms to tackle the lack of QE training data for related languages. We show results on a small sample for the English to Romance languages directions, and we test the contribution of related languages also on the en-de test dataset. The STL model shows to outperform the other transfer methods. However, this model is sensible to the balance of the labelled training data, so that a different balance in the unlabelled dataset affects the final performance. We tried to overcome the unbalanced data by adding artificial instances for the under represented class, but this heuristic was not successful.

For future work, we plan to extend the testing with various annotators in order to acquire reasonable testing datasets for the language pairs under study. We will add extra features to the QuEst baseline based on similarity scores as domain indicators to characterise differences and similarities between domains. We will also expand the available labelled resources into other language families given that the STL only requires as input a small amount of labelled data and larger amounts of unlabelled data, where we can expand QE across related languages. Finally, we would like to try converting the QE models for translation **into** related languages to a model for estimating the translation quality **between** these languages, for example, using en-es and en-pt models to estimate the quality of es-pt translations.

Acknowledgments

The research was funded by Innovate UK and ZOO Digital Group plc.
`http://www.zoodigital.com/`

References

Eleftherios Avramidis. 2014. Efforts on machine learning over human-mediated translation edit rate. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 302–306, Baltimore, Maryland, USA, June.

Daniel Beck, Kashif Shah, and Lucia Specia. 2014. Shef-lite 2.0: Sparse multi-task gaussian processes for translation quality estimation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 307–312, Baltimore, Maryland, USA, June.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, February.

Ergun Bicici and Andy Way. 2014. Referential translation machines for predicting translation quality. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 313–321, Baltimore, Maryland, USA, June.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting Success in Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii.

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 Guilherme Camargo de Souza, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014. 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.

José Guilherme Camargo de Souza, Matteo Negri, Elisa Ricci, and Marco Turchi. 2015. Online multitask learning for machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 219–228.

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, Stroudsburg, PA, USA.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 759–766, New York, NY, USA.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proc 13th Conference of the European Association for Machine Translation*, pages 28–37.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. Quest - a translation quality estimation framework. In *51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL, pages 79–84, Sofia, Bulgaria.

Marco Turchi and Matteo Negri. 2014. Automatic annotation of machine translation datasets with binary quality judgements. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Yi Yang and Jacob Eisenstein. 2015. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 672–682, Denver, Colorado, May–June.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. In *online*.