Machine Translation Quality Estimation Across Domains

José G. C. de Souza
University of Trento
Fondazione Bruno Kessler
Trento, Italy
desouza@fbk.eu

Marco Turchi
Fondazione Bruno Kessler
Trento, Italy
turchi@fbk.eu

Matteo Negri
Fondazione Bruno Kessler
Trento, Italy
negri@fbk.eu

Abstract

Machine Translation (MT) Quality Estimation (QE) aims to automatically measure the quality of MT system output without reference translations. In spite of the progress achieved in recent years, current MT QE systems are not capable of dealing with data coming from different train/test distributions or domains, and scenarios in which training data is scarce. We investigate different multitask learning methods that can cope with such limitations and show that they overcome current state-of-the-art methods in real-world conditions where training and test data come from different domains.

1 Introduction

Machine Translation (MT) Quality Estimation (QE) aims to automatically predict the quality of MT output without using reference translations (Blatz et al., 2003; Specia et al., 2009). QE systems usually employ supervised machine learning models that use different information extracted from (source, target) sentence pairs as features along with quality scores as labels. The notion of quality that these models measure can be indicated by different scores. Some examples are the average number of edits required to post-edit the MT output, i.e., human translation edit rate[1] (HTER (Snover et al., 2006)), and the time (in seconds) required to post-edit a translation produced by an MT system (Specia, 2011).

Research on QE has received a strong boost in recent years due to the increase in the usage of MT systems in real-world applications. Automatic and reference-free MT quality prediction demonstrated to be useful for different applications, such as: deciding whether the translation output can be published without post-editing (Soricut and Echihabi, 2010), filtering out low-quality translation suggestions that should be rewritten from scratch (Specia et al., 2009), selecting the best translation output from a pool of MT systems (Specia et al., 2010), and informing readers of the translation whether it is reliable or not (Turchi et al., 2012). Another example is the computer-assisted translation (CAT) scenario, in which it might be necessary to predict the quality of translation suggestions generated by different MT systems to support the activity of post editors working with different genres of text.

The dominant QE framework presents some characteristics that can limit models' applicability in such 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 labeled data. Second, the learning methods currently used (for instance in the framework of QE shared evaluation campaigns)[2] 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 as different data in real-world applications might be post-edited by different translators, the translations might be generated by different MT systems and the documents being translated might belong to different domains or genres. To

[1] 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.

[2] In the last two editions of the yearly Workshop on Machine Translation, several QE shared tasks have been proposed (Callison-Burch et al., 2012; Bojar et al., 2013).

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 409–420, Dublin, Ireland, August 23-29 2014.

overcome these limitations a plausible research objective is to exploit techniques that: (i) allow domains and distributions of features to be different between training and test data, and (ii) that cope with the scarce amount of training labels by sharing information across domains, a common scenario for transfer learning.

In this paper we investigate the use of techniques that can exploit the training instances from different domains to learn a QE model for a specific target domain for which there is a small amount of labeled data. In particular, we are interested in approaches that allow not only learning from one single source domain but also from multiple source domains simultaneously, by leveraging the labels from all available data to improve results in a target domain.

Given these requirements, we experiment with different *multitask learning* techniques that perform transfer learning via a common task structure (domain relatedness). Furthermore, we employ an approach based on *feature augmentation* that has been successfully used in other natural language processing tasks. We present a series of experiments over three domains with increasing amounts of training data, showing that our adaptive approaches outperform competitive baselines.

The contributions of our work are: (i) a first exploration of techniques that overcome the limitation of current QE learning methods when dealing with data with different training and test distributions and domains, and (ii) an empirical verification of the amount of training data required by such techniques to outperform competitive baselines on different target domains. To the best of our knowledge, this is the first work addressing the challenges posed by domain adaptation in MT QE.

2 Related Work

Quality estimation has recently gained increasing attention, also boosted by two evaluation campaigns organized within the Workshop on Machine Translation (WMT) (Callison-Burch et al., 2012; Bojar et al., 2013). The bulk of work done so far has focused on the controlled WMT evaluation framework and, in particular, on two major aspects of the problem: feature engineering and machine learning methods.

Feature engineering accounts for linguistically-based predictors that aim to model different perspectives of the quality estimation problem. The research ranges from identifying indicators that approximate the complexity of translating the source sentence and designing features that model the fluency of the automatically generated translation, to linguistically motivated measures that estimate how adequate the translation is in comparison to the source sentence in terms of meaning (Blatz et al., 2003; Mehdad et al., 2012; Hardmeier et al., 2012; Rubino et al., 2012; Specia et al., 2012; de Souza et al., 2013a).

State-of-the-art QE explores different supervised linear or non-linear learning methods for regression or classification such as Support Vector Machines (SVM), different types of Decision Trees, Neural Networks, Elastic-Net, Gaussian Processes, Naive Bayes, among others (Specia et al., 2009; Buck, 2012; Beck et al., 2013; Souza et al., 2014). Another aspect related to the learning methods that has received attention is the optimal selection of features in order to overcome issues related with the high-dimensionality of the feature space (Soricut et al., 2012; de Souza et al., 2013a; Beck et al., 2013; de Souza et al., 2013b).

Despite constant improvements, such learning methods have limitations. The main one is that they assume that both training and test data are independently and identically distributed. As a consequence, when they are applied to data from a different distribution or domain they show poor performance. This limitation harms the performance of QE systems for several real-world applications, such as CAT environments. Advanced CAT systems currently integrate suggestions obtained from MT engines with those derived from translation memories (TMs). In such framework, the compelling need to speed up the translation process and reduce its costs by presenting human translators with good-quality suggestions raises interesting research challenges for the QE community. In such environments, translation jobs come from different domains that might be translated by different MT systems and are routed to professional translators with different idiolect, background and quality standards (Turchi et al., 2013). Such variability calls for flexible and adaptive QE solutions by investigating two directions: (i) modeling translator behaviour (Turchi et al., 2014) and (ii) maximize the learning capabilities from all the available data. The second research objective motivates our investigation on methods that allow the training and test domains and

the distributions to be different.

Recent work in QE focused on aspects that are problematic even in the controlled WMT scenario, and are closely related to the flexibility/adaptability issue. Focusing on the first of the two aforementioned directions (i.e. modeling translators' behaviour), Cohn and Specia (2013) propose a Multitask Gaussian Process method that jointly learns a series of annotator-specific models and that outperforms models trained for each annotator. Our work differs from theirs in that we are interested in the latter research direction (i.e. coping with domain and distribution diversity) and we use in and out-of-domain data to learn robust in-domain models. Our scenario represents a more challenging setting than the one tackled in (Cohn and Specia, 2013), which does not consider different domains.

In *transfer learning* there are many techniques suitable to fulfill our requirements. The aim of transfer learning is to extract the knowledge from one or more source tasks and apply it to a target task (Pan and Yang, 2010). One type of transfer learning is *multitask learning* (MTL), which 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).

Domain adaptation is a kind of transfer learning in which source and target domains (i.e. training and test) are different but the tasks are the same (Pan and Yang, 2010). The domain adaptation techniques that inspire our work have been successfully applied to a variety of NLP tasks (Blitzer et al., 2006; Jiang and Zhai, 2007). For instance, an effective solution for supervised domain adaptation, EasyAdapt (SVR FEDA henceforth), was proposed in (Daumé III, 2007) and applied to named entity recognition, part-of-speech tagging and shallow parsing. The approach transforms the domain adaptation problem into 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 one has access to out-of-domain labels and wants to leverage a small amount of available in-domain labeled data to train a model (Daumé III, 2007), the case of this study. This is different from the *semi-supervised* case in which in-domain labels are not available.

3 Adaptation for QE

An important assumption in MTL is that different tasks (domains in our case) are correlated via a certain structure. Examples of such structures are the hidden layers in a neural network (Caruana, 1997) and shared feature representation (Argyriou et al., 2007) among others. This common structure allows for knowledge transfer among tasks and has been demonstrated to improve model generalization over single task learning (STL) for different problems in different areas. Under this scenario, several assumptions can be made about the relatedness among the tasks, leading to different transfer structures. We explore three approaches to MTL that deal with task relatedness in different ways. These are the "Dirty" approach to MTL (Jalali et al., 2010), Sparse Trace MTL (Chen et al., 2012) and Robust MTL (Chen et al., 2011). The three approaches use different regularization techniques that capture task relatedness using norms over the weights of the features.

Before describing the three approaches, we introduce some basic notation similar to (Chen et al., 2011). In MTL there are T tasks and each task $t \in T$ has m training samples $\{(x_1^{(t)}, y_1^{(t)}), \ldots, (x_m^{(t)}, y_m^{(t)})\}$, with $x_i^{(t)} \in \mathbb{R}^d$ where d is the number of features and $y_i^{(t)} \in \mathbb{R}$ is the output (the response variable or label). The input features and labels are stacked together to form two different matrices $X^{(t)} = [x_1^{(t)}, \ldots, x_m^{(t)}]$ and $Y^{(t)} = [x_1^{(t)}, \ldots, x_m^{(t)}]$, respectively. The weights of the features for each task are represented by W, where each column corresponds to a task and each row corresponds to a feature.

The **"Dirty"** approach to MTL follows the idea that different tasks may share the same discriminative features (Argyriou et al., 2007). However, it also considers that different tasks might have different discriminative features that are inherent to each task. Therefore, the method encourages shared-sparsity among tasks and among features in each task. It decomposes W into two components, one is a row-

sparsed matrix that corresponds to the features shared among the tasks and the other is an element-wise sparse matrix that corresponds to the non-shared features that are important for each task independently. More formally, the "Dirty" approach is explained by Equation 1.

$$\min_{W} \sum_{t=1}^{T} ||(W^{(t)}X^{(t)} - Y^{(t)})||_2^2 + \lambda_s||S||_1 + \lambda_b||B||_{1,\infty} \text{ subject to: } W = S + B \quad (1)$$

where $||(W^{(t)}X^{(t)} - Y^{(t)})||_2^2$ is the least squares loss function, S is the regularization term that encourages element-wise sparsity and B is the block-structured row-sparsity regularizer. The $||.||_2$ is the l_2-norm (Euclidean distance), $||.||_1$ is the l_1-norm (given by $\sum_{i=1} |x_i|$) and $||.||_{1,\infty}$ is the row grouped l_1-norm. The λ_s and λ_b are non-negative trade-off parameters that control the amount of regularization applied to S and B, respectively.

Sparse Trace MTL considers the problem of learning incoherent sparse and low-rank patterns from multiple related tasks. This approach captures task relationship via a shared low-rank structure of the weight matrix W. As computing the low-rank structure of a matrix leads to a NP-hard optimization problem, Chen et al. (2012) proposed to compute the trace norm as a surrogate, making the optimization problem tractable. In addition to learning the low-rank patterns, this method also considers the fact that different tasks may have different inherent discriminative features. It decomposes W into two components: S, which models element-wise sparsity, and Q, which captures task relationship via the trace norm. The convex problem minimized by Sparse Trace is given in Equation 2.

$$\min_{W} \sum_{t=1}^{T} ||(W^{(t)}X^{(t)} - Y^{(t)})||_2^2 + \lambda_s||S||_1 \text{ subject to: } W = S + Q, ||Q||_* < \lambda_p \quad (2)$$

where $||.||_*$ is the trace norm, given by the sum of the singular values σ_i of W, i.e., $||W||_* = \sum_{i=1} \sigma_i(W)$. Here, λ_p controls the rank of Q and λ_s controls the sparsity of S.

The key assumption in MTL is that tasks are related in some way. However, this assumption might not hold for a series of real-world problems. In situations in which tasks are not related a negative transfer of information among tasks might occur, harming the generalization of the model. One way to deal with this problem is to: (i) group related tasks in one structure and share knowledge among them, and (ii) identify irrelevant tasks maintaining them in a different group that does not share information with the first group. This is the idea of **Robust** MTL (RMTL henceforth). The algorithm approximates task relatedness via a low-rank structure like Sparse Trace and identifies outlier tasks using a group-sparse structure (column-sparse, at task level). Robust MTL is described by Equation 3. It employs a non-negative linear combination of the trace norm (the task relatedness component L) and a column-sparse structure induced by the $l_{1,2}$-norm (the outlier task detection component S). If a task is an outlier it will have non-zero entries in S.

$$\min_{W} \sum_{t=1}^{T} ||(W^{(t)}X^{(t)} - Y^{(t)})||_2^2 + \lambda_l||L||_* + \lambda_s||S||_{1,2} \text{ subject to: } W = L + S \quad (3)$$

where $||S||_{1,2}$ is the group regularizer that induces sparsity on the tasks.

4 Experimental Setting

In this section we describe the data used for our experiments, the features extracted, the set up of the learning methods, the baselines used for comparison and the evaluation of the models. The goal of our experiments is to show that the methods presented in Section 3 outperform competitive baselines and standard QE learning methods that are not capable of adapting to different domains. We experiment with three different domains of comparable size and evaluate the performance of the adaptive methods and the standard techniques with different amounts of training data. The MTL models described in section 3 are trained with the Malsar toolkit implementation (Zhou et al., 2012). The hyper-parameters are optimized

using 5-fold cross-validation in a grid search procedure. The parameter values are searched in an interval ranging from 10^{-3} to 10^{3}.

4.1 Data

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 are 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 use as labels the HTER score computed with TERCpp[3]. For the three domains we use half of the data for training (181 instances) and half of the data for testing (182 instances). The limited 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 and representative training sets is far from being guaranteed (e.g. the CAT scenario).

The sentence tuples for the first two domains are randomly sampled from the Trace corpus[4]. The translations were generated by two different MT systems, a state-of-the-art phrase-based statistical MT system and a commercial rule-based system. Furthermore, the translations were post-edited by up to four different translators, as described in (Wisniewski et al., 2013).

Domain	No. of tokens	Vocab. size	Avg. sent. length
TED source	6858	1659	19
TED target	7016	1828	19
IT source	3310	1004	9
IT target	3134	1049	8
News source	7605	2273	21
News target	8230	2346	23

Table 1: Datasets statistics for each domain.

The TED talks domain is formed by subtitles of several talks in a range of topics presented in the TED conferences. The complete dataset has been used for MT and automatic speech recognition systems evaluation within the International Workshop on Spoken Language Translation (IWSLT). The News domain is formed by newswire text used in WMT translation campaigns and covers different topics. 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 on the Matecat[5] CAT tool in real working conditions. Table 1 provides macro-indicators (number of tokens, vocabulary size, average sentence length) that evidence the large difference between the domains addressed by our experiments and give an idea of the difficulty of the task.

A peculiarity of the TED domain is that it is formed by manual transcriptions of speech translated by different MT systems, configuring a different type of discourse than News and IT. In TED, the vocabulary size in the source and target sentences is lower than that of the News domain but higher than IT. News presents the most varied vocabulary, which is an evidence of the more varied lexical choice represented by the several topics that compose the domain. Moreover, News has the highest average sentence length, a characteristic of non-technical written discourse, which tends to have longer sentences than spoken discourse and domains dominated by technical jargon. Such a characteristic is exactly what differentiates IT from the other two domains. IT sentences are technical and present a reduced average number of

[3] http://sourceforge.net/projects/tercpp/

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

[5] www.matecat.com

words, as evidenced by the vocabulary size (the smallest among the three domains). These numbers suggest a divergence between IT and the other two domains, possibly making adaptation more difficult.

4.2 Features

For all the experiments we use the same feature set composed of seventeen features proposed in (Specia et al., 2009). The set is formed by features that model the complexity of translating the source sentence (e.g. the average source token length or the number of tokens in the source sentence), and the fluency of the translated sentence produced by the MT system (e.g. the language model probability of the translation). The decision to use this feature set is motivated by the fact that it demonstrated to be robust across language pairs, MT systems and text domains (Specia et al., 2009). The 17 features are:

- number of tokens in the source sentence and in the generated translation;
- average source token length;
- average number of occurences of the target word within the generated translation;
- language model probability of the source sentence and generated translation;
- average number of translations per source word in the sentence: as given by IBM 1 model thresholded so that $P(t|s) > 0.2$ weighted by the inverse frequency of each word in the source side of the SMT training corpus ◯;
- average number of translations per source word in the sentence: as given by IBM 1 model thresholded so that $P(t|s) > 0.01$ weighted by the inverse frequency of each word in the source side of the SMT training corpus;
- percentage of unigrams◯, bigrams and trigrams◯ in the first quartile of frequency (lower frequency words) in a corpus of the source language;
- percentage of unigrams◯, bigrams and trigrams in the fourth quartile of frequency (higher frequency words) in a corpus of the source language;
- percentage of unigrams in the source sentence seen in the source side of the SMT training corpus;
- number of punctuation marks in the source sentence and in the hypothesis translation;

4.3 Baselines

As a term of comparison, we consider these baselines in our experiments. A simple to implement but difficult to beat baseline when dealing with regression on tasks with different distributions is to compute the mean of the training labels and use it as the prediction for each testing point (Rubino et al., 2013). Hereafter we refer to this baseline as μ. Since supervised domain adaptation techniques should outperform models that are trained only on the available in-domain data, we also use as baseline the regressor built only on the available in-domain data (SVR in-domain). Furthermore, as a third baseline, we train a regressor by pooling together training data of all domains, combining source and target data without any kind of task relationship mechanism (SVR Pooling).

The baselines are trained on the feature set described earlier in Section 4.2 with an SVM regression (SVR) method using the implementation of Scikit-learn (Pedregosa et al., 2011). The radial basis function (RBF) kernel is used for all baselines. The hyper-parameters of the model are optimized using randomized search optimization process with 50 iterations as described in (Bergstra and Bengio, 2012) and used previously for QE in (de Souza et al., 2013a). The best parameters are found using 5-fold cross-validation on the training data and ϵ, γ and C are sampled from exponential distributions scaled at 0.1 for the first two parameters and scaled at 100 for the last one. It is important to notice that the SVR with RBF kernel methods learn non-linear models that have been shown to perform better than linear models on the set of features used for predicting HTER. On the contrary, the MTL methods presented in Section 3 are methods that do not explore kernels or any other kind of non-linear learning method.

Source / Target	IT_{tgt}	$News_{tgt}$	TED_{tgt}
IT_{src}	0.2081	0.2341	0.2232
$News_{src}$	0.2368	0.1690	0.2130
TED_{src}	0.2183	0.2263	0.1928

Table 2: Results of the SVR in-domain baseline trained and evaluated in each domain (average of 50 different shuffles). Rows represent the domain data used to train the model and columns represent the domain data used to evaluate the model. Scores are MAE.

4.4 Evaluation

The accuracy of the models is evaluated with the mean absolute error (MAE), which was also used in previous work and in the WMT QE shared tasks (Bojar et al., 2013). MAE is the average of the absolute difference between the prediction $\hat{y}_i$ of a model and the gold standard response y_i (Equation 4). As it is an error measure, lower values mean better performance.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i - y_i| \tag{4}$$

To test the statistical significance of our results we need to perform comparisons of multiple models. In addition, we would like to test the significance over different training amounts. Given these requirements we need to perform multiple hypothesis tests instead of paired tests. It has been shown that for comparisons of multiple machine learning models, the recommended approach is to use a non-parametric multiple hypothesis test followed by a post-hoc analysis that compares each pair of hypothesis (Demšar, 2006). In our experiments we use the Friedman test (Friedman, 1937; Friedman, 1940) followed by a post-hoc analysis of the pairs of regressors using Holm's procedure (Holm, 1979) to perform the pairwise comparisons when the null hypothesis is rejected. All tests for both Friedman and post-hoc analysis are run with $\alpha = 0.05$. For more details about these methods, we refer the reader to (Demšar, 2006; Garcia and Herrera, 2008) which provide a complete review about the application of multiple hypothesis testing to machine learning methods.

5 Results and Discussion

Our experiments are organized as follows. First, we evaluate the performance of single task learning methods on different cross-domain experiments. Then, we report the evaluation for the multitask learning methods and discuss the results.

5.1 Single Task Learning

With the objective of having an insight about the difference between the domains, we train the SVR in-domain baseline with all available training data for each domain and evaluate its performance on the same domain and in the two remaining domains.

Results are reported in Table 2, where the diagonal shows the figures for the in-domain evaluation. These numbers suggest that the IT domain configures a more difficult challenge for the learning algorithm. The IT in-domain model (IT_{src}-IT_{tgt}) presents a performance 21% inferior to News and 8% inferior to TED. For all models trained on a source domain different than the target domain there is a drop in performance, as it is expected from a system that assumes that training and test data are sampled from the same distribution. In addition, when predicting IT using the model trained on News, we have a perfomance drop of 13% whereas using the model trained on TED the performance drops up to 4%.

5.2 Multitask learning

We run the baselines described in Section 4.3 and the methods described in Section 3 on different amounts of training data, ranging from 18 to 181 instances (10% and 100%, respectively). The motivation is to verify how much training data is required by the MTL methods to outperform the baselines for a target domain. Table 3 presents the results for the three domains with models trained on 30, 50 and

100% of the training data (54, 90 and 181 instances, respectively). Each method was run on 50 different train/test splits of the data in order to account for the variability of points in each split.

Method	TED	News	IT
30 % of training data (54 instances)			
mean	0.1951	0.1711	0.2174
SVR In-Domain	0.2013	0.1753	0.2235
SVR Pooling	0.1962	0.1899	0.2201
SVR FEDA	0.1952	0.1839	0.2193
MTL Dirty	0.1954	0.1708	0.2193
MTL SparseTrace	0.1976	0.1743	0.2222
MTL RMTL	**0.1946**	**0.1685**	**0.2162**
50% of training data (90 instances)			
mean	0.1943	0.1707	0.2170
SVR In-Domain	0.1976	0.1711	0.2183
SVR Pooling	0.1951	0.1865	0.2191
SVR FEDA	0.1937	0.1806	0.2161
MTL Dirty	0.1927	0.1678	0.2148
MTL SparseTrace	0.1922	0.1672	0.2157
MTL RMTL	**0.1878**	**0.1653**	**0.2119**
100% of training data (181 instances)			
mean	0.1936	0.1690	0.2162
SVR In-Domain	0.1928	0.1690	0.2081
SVR Pooling	0.1927	0.1849	0.2203
SVR FEDA	0.1908	0.1757	0.2107
MTL Dirty	0.1878	0.1666	0.2083
MTL SparseTrace	0.1881	0.1661	0.2094
MTL RMTL	**0.1846**	**0.1653**	**0.2075**

Table 3: Average performance of fifty runs of the models on different train and test splits with 30, 50 and 100 percent of training data. The average scores reported are the MAE.

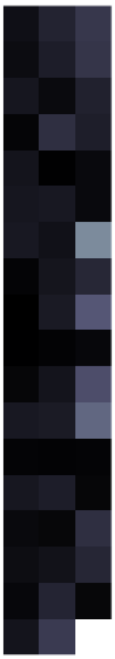

Figure 1: Visualization of the RMTL task outlier model when trained on all the 181 instances of training data. Cells with darker shades are closer to zero. Cells with lighter shades are closer to one. Columns with only black entries are considered inlier tasks (domains). From left to right, columns correspond to News, TED and IT domains. The first 17 rows correspond to the features used to train the model and the last row in corresponds to the bias term.

For all three domains, a general trend is that MTL RMTL is the method that reaches the lowest MAE when compared to all the other models. Given the difference among the domains, it is very likely that MTL Dirty and MTL SparseTrace suffer from the negative transfer problem (the assumption that all tasks are similar does not hold). MTL RMTL is the only method among the methods presented here that copes with negative transfer among tasks. The significance tests indicate that MTL RMTL improvements are statistically significant with respect to all baselines depending on the range of training data used to compute the test.

- For **TED**, the Friedman test rejects the null hypothesis with $p = 4.62^{-5}$. Post-hoc analysis indicates that there are differences statistically significant between MTL RMTL and all the three baselines with $p \leq 0.002$.

- For **News**, the Friedman test measures significant differences with $p = 1.14^{-9}$ and the post-hoc analysis indicates that MTL RMTL is statistically significant with respect to SVR in-domain and SVR Pooling with $p = 0.002$ for varying amounts of training data from 10 to 100%. As can be seen in Figure 2, MTL RMTL starts with a very high MAE using 10% of the data (approximately 0.21 MAE) but improves dramatically with 20% of the data. Calculating the significance test with 20 to 100% of training data, MTL RMTL is significantly better than all baselines with $p \leq 2.89^{-10}$.

- For **IT**, in a similar situation to the News domain, RMTL is significantly better than all baselines

trained on 30% to 100% of the training data (Friedman test's $p = 2.86^{-4}$ and post-hoc analysis' $p \leq 3.73^{-7}$).

Another observed trend is that the MTL models benefit from increasing amounts of training data. MTL RMTL has an improvement in performance of 5.13% for TED, 4% for News and 1.85% for IT when trained on 100% of the training data in comparison with the model trained on 30% of training data.

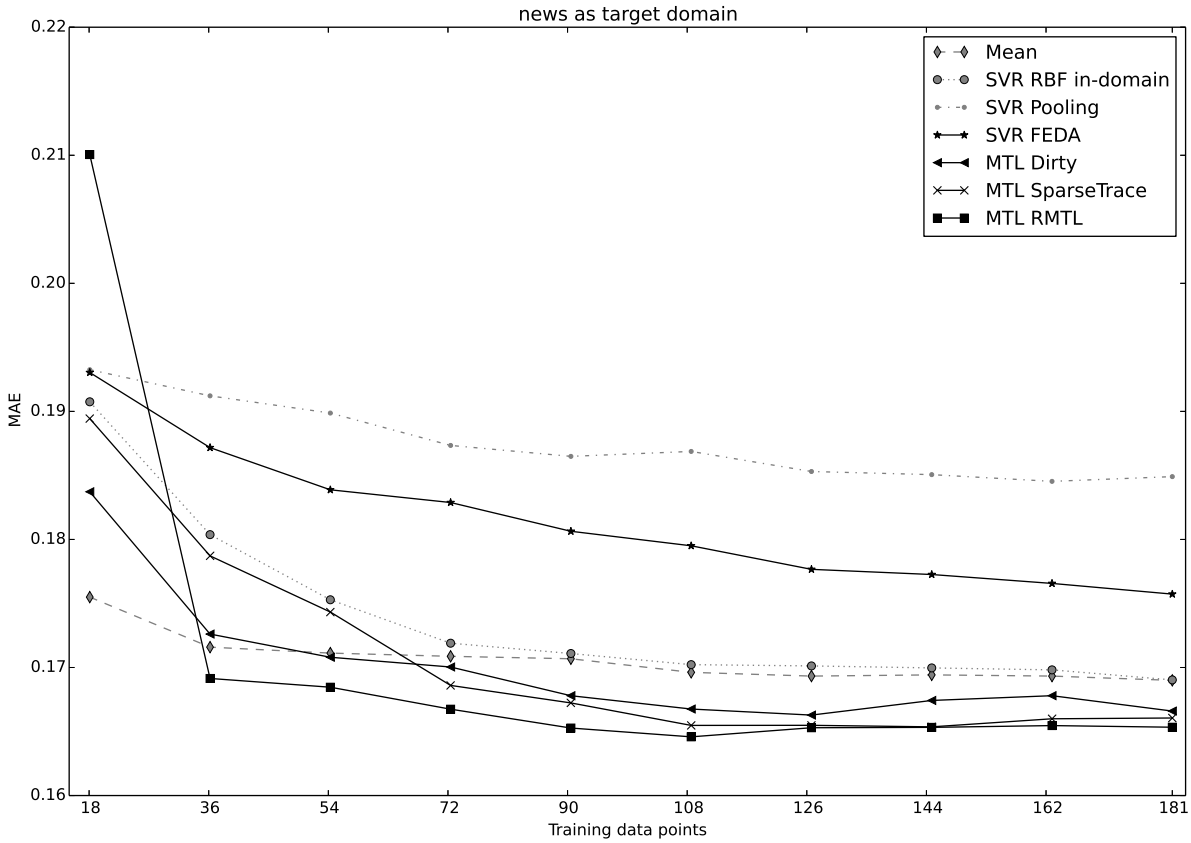

Figure 2: Learning curves for the News domain.

The results for the IT domain are in line with the in-domain experiments in which we observed that IT is a more challenging domain in comparison to TED and News. The MAE of IT is always higher than for the other domains on in-domain and MTL experiments. Another evidence of this is the model learned by the RMTL method when using all training data and run on one of the 50 training/test splits. A graphic representation of the RMTL outlier task detection component (described in Section 3) is shown in Figure 1.

From left to right, each column represents News, TED, and IT domains, respectively, while each row is the instantiation of a feature in the corresponding task. Columns with non-black entries represent outlier tasks. The highest number of entries with lighter shades is in the third column, IT. Several features in this task are considered outliers with respect to the same features in the other tasks. Consequently, the learning method takes the weights into consideration to a greater extent when learned with the outlier model for the IT domain. Entries with the lightest shades in the IT domain correspond to the features marked with ◯ in Section 4.2. These outlier features are directly affected by the length of the sentences on which they are computed (source or target) given that the number of tokens influences the final value of the feature. This outcome goes in the same direction of our analysis of the three domains (Section 4.1) that indicates a very different vocabulary size and average sentence length for IT when compared to the other two domains.

To a lesser extent than IT, News and TED domains also present a few lighter-shaded entries in the outlier component (1st and 2nd column). This suggests that MTL RMTL was capable of transfering information among the domains in a more efficient way than the other MTL methods analyzed.

Overall the experiments presented show encouraging results in the direction of coping with QE data coming from different domains/genres, translated by different MT systems and post-edited by different translators. Results show that even in such difficult conditions, the methods investigated are capable of outperforming competitive baselines based on non-linear models on different domains. As a rationale, models that consider not only similarity between the domains but also deal with some sort of dissimilarity should be considered. This is the case of the best performing method, MTL RMTL, which identifies outlier tasks in order to avoid negative transfer among tasks.

6 Conclusion

In this work we presented an investigation of methods that overcome limitations presented by current MT QE state-of-the-art systems when applied to real world conditions. In such scenarios (e.g. CAT environment) the requirements are two-fold: (i) learning in the presence of different train/test feature and label distributions and across different domains/genres, and (ii) the capability of learning with scarce training data. In our experiments, we explored transfer learning methods, in particular multitask learning, and we showed that such methods can cope with the needs of real-world scenarios.

We showed that multitask learning methods are capable to learn robust models for three different domains that perform better than three strong baselines trained on the same amount of data. The methods explored here benefit from increasing amounts of training data but also perform well when operating with very limited amounts of data. We believe that the results obtained in this first exploration of model adaptation for the problem can encourage the MT QE community to shift the focus from controlled scenarios to more applicable, real-world contexts that require more robust methods.

Acknowledgements

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

References

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2007. Multi-task feature learning. In *Advances in neural information processing systems*, volume 19.

Daniel Beck, Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. SHEF-Lite: When less is more for translation quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 337–342.

James Bergstra and Yoshua Bengio. 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305.

Joseph John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. In *20th COLING*, pages 315–321.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Morristown, NJ, USA. Association for Computational Linguistics.

Ondej 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 *Eighth Workshop on Statistical Machine Translation*, pages 1–44.

Christian Buck. 2012. Black Box Features for the WMT 2012 Quality Estimation Shared Task. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 91–95.

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*, pages 10–51, Montr{é}al, Canada, June. Association for Computational Linguistics.

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.

Jianhui Chen, Ji Liu, and Jieping Ye. 2012. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Transactions on Knowledge Discovery from Data*, 5(4):22, February.

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*, pages 32–42.

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

José G. C. de Souza, Christian Buck, Marco Turchi, and Matteo Negri. 2013a. FBK-UEdin participation to the WMT13 Quality Estimation shared-task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 352–358.

José G.C. de Souza, Miquel Esplà-Gomis, Marco Turchi, and Matteo Negri. 2013b. Exploiting qualitative information from automatic word alignment for cross-lingual nlp tasks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 771–776, Sofia, Bulgaria, August. Association for Computational Linguistics.

Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, 7:1–30, December.

Milton Friedman. 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701.

Milton Friedman. 1940. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

Salvador Garcia and Francisco Herrera. 2008. An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research*, 9:2677–2694.

Christian Hardmeier, Joakim Nivre, and Jorg Tiedemann. 2012. Tree Kernels for Machine Translation Quality Estimation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, number 2011, pages 109–113.

Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):pp. 65–70.

Ali Jalali, PD Ravikumar, S Sanghavi, and C Ruan. 2010. A Dirty Model for Multi-task Learning. In *Advances in Neural Information Processing Systems (NIPS) 23*.

Jing Jiang and Chengxiang Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, number June, pages 264–271.

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 Zenz, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007 Demo and Poster Sessions*, number June, 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.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Mathieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Raphael Rubino, Jennifer Foster, Joachim Wagner, Johann Roturier, Rasul Samad Zadeh Kaljahi, and Fred Hollowood. 2012. DCU-Symantec Submission for the WMT 2012 Quality Estimation Task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 138–144, Montr{é}al, Canada, June. Association for Computational Linguistics.

Raphael Rubino, José G. C. de Souza, Jennifer Foster, and Lucia Specia. 2013. Topic Models for Translation Quality Estimation for Gisting Purposes. In *Machine Translation Summit (MT Summit) XIV*, pages 295–302.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Association for Machine Translation in the Americas*.

Radu Soricut and A Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, number July, pages 612–621.

Radu Soricut, Nguyen Bach, and Ziyuan Wang. 2012. The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 145–151.

José G. C. 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*, Baltimore, MD, USA, June.

Lucia Specia, Marco Turchi, Nello Cristianini, Nicola Cancedda, and Marc Dymetman. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the EAMT*, number May, pages 28–35.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50, May.

Lucia Specia, Stafford Street, Regent Court, and Mariano Felice. 2012. Linguistic Features for Quality Estimation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 96–103.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the European Association for Machine Translation*, number May, pages 73–80.

Marco Turchi, Josef Steinberger, and Lucia Specia. 2012. Relevance ranking for translated texts. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, number May, pages 153–160.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proceedings of the 8th Workshop on Statistical Machine Translation (WMT'13)*, Sofia, Bulgaria, August.

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*.

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.

Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2012. MALSAR: Multi-tAsk Learning via StructurAl Regularization.