Large-scale Multitask Learning for Machine Translation Quality Estimation

Kashif Shah and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{kashif.shah,l.specia}@sheffield.ac.uk

Abstract

Multitask learning has been proven a useful technique in a number of Natural Language Processing applications where data is scarce and naturally diverse. Examples include learning from data of different domains and learning from labels provided by multiple annotators. *Tasks* in these scenarios would be the domains or the annotators. When faced with limited data for each task, a framework for the learning of tasks in parallel while using a shared representation is clearly helpful: what is learned for a given task can be transferred to other tasks while the peculiarities of each task are still modelled. Focusing on machine translation quality estimation as application, in this paper we show that multitask learning is also useful in cases where data is abundant. Based on two large-scale datasets, we explore models with multiple annotators and multiple languages and show that state-of-the-art multitask learning algorithms lead to improved results in all settings.

1 Introduction

Quality Estimation (QE) models predict the quality of Machine Translation (MT) output based on the source and target texts only, without reference translations. This task is often framed as a supervised machine learning problem using various features indicating fluency, adequacy and complexity of the source-target text pair, and annotations on translation quality given by human translators. Various kernel-based regression and classification algorithms have been explored to learn prediction models.

The application of QE we focus on here is that of guiding professional translators during the post-editing of MT output. QE models can provide translators with information on how much editing/time will be necessary to fix a given segment, or on whether it is worth editing it at all, as opposed to translating it from scratch. For this application, models are learnt from quality annotations that reflect post-editing effort, for instance, 1-5 judgements on estimated post-editing effort (Callison-Burch et al., 2012) or actual post-editing effort measured as post-editing time (Bojar et al., 2013) or edit distance between the MT output and its post-edited version (Bojar et al., 2014; Bojar et al., 2015).

One of the biggest challenges in this field is to deal with the inherent subjectivity of quality labels given by humans. Explicit judgements (e.g. the 1-5 point scale) are affected the most, with previous work showing that translators' perception of post-editing effort differs from actual effort (Koponen, 2012). However, even objective annotations of actual post-editing effort are subject to natural variance. Take, for example, post-editing time as a label: Different annotators have different typing speeds and may require more or less time to deal with the same edits depending on their level of experience, familiarity with the domain, etc. Post-editing distance also varies across translators as there are often multiple ways of producing a good quality translation from an MT output, even when strict guidelines are given.

In order to address variance among multiple translators, three strategies have been applied: (i) models are built by averaging annotations from multiple

Proceedings of NAACL-HLT 2016, pages 558–567,
San Diego, California, June 12-17, 2016.

translators on the same data points, as was done in the first shared task on the topic (Callison-Burch et al., 2012); (ii) models are built for individual translators by collecting labelled data for each translator (Shah and Specia, 2014); and (iii) models are built using multitask learning techniques (Caruana, 1997) to put together annotations from multiple translators while keeping track of the translators' identification to account for their individual biases (Cohn and Specia, 2013; de Souza et al., 2015).

The first approach is sensible because, in the limit, the models built should reflect the "average" strategies/preferences of translators. However, its cost makes it prohibitive. The second approach can lead to very accurate models but it requires sufficient training data for each translator, and that all translators are known at model building time. The last approach is very attractive. It is a *transfer learning* (a.k.a. *domain-adaptation*) approach that allows the modelling of data from each individual translator while also modelling correlations between translators such that "similar" translators can mutually inform one another. As such, it does not require multiple annotations of the same data points and can be effective even if only a few data points are available for each translator. In fact, previous work on multitask learning for quality estimation has concentrated on the problem of learning prediction models from little data provided by different annotators.

In this paper we take a step further to investigate multitask learning for quality estimation in settings where data may be abundant for some or most annotators. We explore a multitask learning approach that provides a general, scalable and robust solution regardless of the amount of data available. By testing models on single translator data, we show that while building models for individual translators is a sensible decision when large amounts of data are available, the multitask learning approach can outperform these models by learning from data by multiple annotators. Additionally, besides having translators as "tasks", we address the problem of learning from data for multiple language pairs.

We devise our multitaslk approach within the Bayesian non-parametric machine learning framework of Gaussian Processes (Rasmussen and Williams, 2006). Gaussian Processes have shown very good results for quality estimation in previous work (Cohn and Specia, 2013; Beck et al., 2013; Shah et al., 2013). Our datasets – annotated for post-editing distance – contain nearly 100K data points, two orders of magnitude larger than those used in previous work. To cope with scalability issues resulting from the size of these datasets, we apply a sparse version of Gaussian Processes. We perform extensive experiments on this large-scale data aiming to answer the following research questions:

- What is the best approach to build models to be used by **individual translators**? How much data is necessary to build independent models (one per translator) that can be as accurate as (or better than) models using data from multiple translators?
- When large amounts of data are available, can we still improve over independent and pooled models by learning from metadata to exploit **transfer across translators**?
- Can crosslingual data help improve model performance by exploiting **transfer across language pairs**?

In the remainder of the paper we start with an overview on related work in the area of multitask learning for quality estimation (Section 2), to then describe our approach to multitask learning in the context of Gaussian Processes (Section 3). In Section 4 we introduce our data and experimental settings. Finally in Sections 5 and 6 we present the results of our experiments to answer the above mentioned questions for cross-annotator and crosslingual transfer, respectively.

2 Related Work

As was discussed in Section 1, the problem of variance among multiple translators in QE has recently been approached in three ways. The first two approaches essentially refer to preparation of the data. At WMT12, the first shared task on QE (Callison-Burch et al., 2012), the official dataset was created by collecting three 1-5 (worst-best) discrete judgements on "perceived" post-editing effort for each translated segment. The final score was a scaled average of the three scores, and about 15% of the labelled data was discarded as annotators diverged in

their judgemetns by more than one point. While this type of data proved useful and certainly reliable in the limit of the number of annotators, it is too expensive to collect.

Shah and Specia (2014) built QE models using data from n annotators by either pooling all the data together or splitting it into n datasets for n individual annotator models. These models were tested in blind versus non-blind settings, where the former refers to test sets whose annotator identifiers were unknown. They observed a substantial difference in the error scores for each of the individual models. They showed that the task is much more challenging for QE models trained independently when training data for each annotator is scarce. In other words, sufficient data needs to be available to build individual models for all possible translators.

The approach of using multitask learning to build models addresses the data scarcity issue and has been shown effective in previous work. Cohn and Specia (2013) first introduced multitask learning for QE. Their goal was to allow the modelling of various perspectives on the data, as given by multiple annotators, while also recognising that they are rarely independent of one another (annotators often agree) by explicitly accounting for inter-annotator correlations. A set of task-specific regression models were built from data labelled with post-editing time and perceived post-editing effort (1-5). "Tasks" included annotators, the MT system and the actual source sentence, as their data included same source segments translated/edited by multiple systems/editors.

Similarly, de Souza et al. studied multitask learning to deal with data coming from different training/test set distributions or domains, and generally scenarios in which training data is scarce. Offline multitask (de Souza et al., 2014a) and online multitask (de Souza et al., 2015; de Souza et al., 2014b) learning methods for QE were proposed. The later focused on continuous model learning and adaptation from new post-edits in a computer-aided translation environment. For that, they adapted an online passive-aggressive algorithm (Cavallanti et al., 2010) to the multitask scenario. While their setting is interesting and could be considered more challenging because of the online adaptation requirements, ours is different as we can take advantage of already having collected large volumes of data.

Multitask learning has also been used for other classification and regression tasks in language processing, mostly for domain adaptation (Daume III, 2007; Finkel and Manning, 2009), but also more recently for tasks such as multi-emotion analysis (Beck et al., 2014), where the each emotion explaining a text is defined as a task. However, in all previous work the focus has been on addressing task variance coupled with data scarcity, which makes them different from the work we describe in this paper.

3 Gaussian Processes

Gaussian Processes (GPs) (Rasmussen and Williams, 2006) are a Bayesian non-parametric machine learning framework considered the state-of-the-art for regression. GPs have been used successfully for MT quality prediction (Cohn and Specia, 2013; Beck et al., 2013; Shah et al., 2013), among other tasks.

GPs assume the presence of a latent function $f : \mathbb{R}^F \rightarrow \mathbb{R}$, which maps a vector $\mathbf{x}$ from feature space F to a scalar value. Formally, this function is drawn from a GP prior:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}')),$$

which is parameterised by a mean function (here, $\mathbf{0}$) and a covariance kernel function $k(\mathbf{x}, \mathbf{x}')$. Each response value is then generated from the function evaluated at the corresponding input, $y_i = f(\mathbf{x}_i)+\eta$, where $\eta \sim \mathcal{N}(0, \sigma_n^2)$ is added white-noise.

Prediction is formulated as a Bayesian inference under the posterior:

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \int_f p(y_*|\mathbf{x}_*, f)p(f|\mathcal{D}),$$

where $\mathbf{x}_*$ is a test input, y_* is the test response value and $\mathcal{D}$ is the training set. The predictive posterior can be solved analitically, resulting in:

$$y_* \sim \mathcal{N}(\mathbf{k}_*^T(\mathbf{K} + \sigma_n^2 I)^{-1}\mathbf{y}, \\ k(x_*, x_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2 I)^{-1}\mathbf{k}_*),$$

where $\mathbf{k}_* = [k(\mathbf{x}_*, \mathbf{x}_1)k(\mathbf{x}_*, \mathbf{x}_2) \dots k(\mathbf{x}_*, \mathbf{x}_n)]^T$ is the vector of kernel evaluations between the training set and the test input and $\mathbf{K}$ is the kernel matrix over the training inputs (the Gram matrix).

3.1 Multitask Learning

The GP regression framework can be extended to multiple outputs by assuming $f(\mathbf{x})$ to be a vector valued function. These models are commonly referred as *Intrinsic Coregionalization Models (ICM)* in the GP literature (Álvarez et al., 2012).

In this work, we employ a separable multitask kernel, similar to the one used by Bonilla et al. (2008) and Cohn and Specia (2013). Considering a set of D tasks, we define the corresponding multitask kernel as:

$$k((\mathbf{x}, d), (\mathbf{x}', d')) = k_{\text{data}}(\mathbf{x}, \mathbf{x}') \times \mathbf{M}_{d,d'},$$

where k_{data} is a kernel (Radial Basis Function, in our experiments) on the input points, d and d' are task or metadata information for each input and $\mathbf{M} \in \mathbb{R}^{D \times D}$ is the multitask matrix, which encodes task covariances. In our experiments, we first consider each post-editor as a different task, and then use crosslingual data to treat each combination of language and post-editor as a task.

An adequate parametrisation of the multitask matrix is required to perform learning process. We follow the parameterisations proposed by Cohn and Specia (2013) and Beck et al. (2014):

Individual: $\mathbf{M} = \mathbf{I}$. In this setting each task is modelled independently by keeping corresponding task identity.

Pooled: $\mathbf{M} = \mathbf{1}$. Here the task identity is ignored. This is equivalent to pooling all datasets in a single task model.

Multitask: $\mathbf{M} = \tilde{\mathbf{H}}\tilde{\mathbf{H}}^T + \text{diag}(\boldsymbol{\alpha})$, where $\tilde{\mathbf{H}}$ is a $D \times R$ matrix. The vector $\boldsymbol{\alpha}$ enables the degree of independence for each task with respect to the global task. The choice of R defines the *rank* ($= 1$ in our case) which can be understood as the capacity of the manifold with which we model the D tasks. We refer readers to see Beck et al. (2014) for a more detailed explanation of this setting.

3.2 Sparse Gaussian Processes

The performance bottleneck for GP models is the Gram matrix inversion, which is $O(n^3)$ for standard GPs, with n being the number of training instances. For multitask settings this becomes an issue for large datasets as the models replicate the instances for each task and the resulting Gram matrix has dimensionality $nd \times nd$, where d is the number of tasks.

Sparse GPs (Snelson and Ghahramani, 2006) tackle this problem by approximating the Gram matrix using only a subset of m inducing inputs. Without loss of generalisation, consider these m points as the first instances in the training data. We can then expand the Gram matrix in the following way:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{mm} & \mathbf{K}_{m(n-m)} \\ \mathbf{K}_{(n-m)m} & \mathbf{K}_{(n-m)(n-m)} \end{bmatrix}.$$

Following the notation in (Rasmussen and Williams, 2006), we refer $\mathbf{K}_{m(n-m)}$ as $\mathbf{K}_{mn}$ and its transpose as $\mathbf{K}_{nm}$. The block structure of $\mathbf{K}$ forms the basis of the so-called Nyström approximation:

$$\tilde{\mathbf{K}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn},$$

which results in the following predictive posterior:

$$\begin{aligned} y_* \sim \mathcal{N}(&\mathbf{k}_{m*}^T \tilde{\mathbf{G}}^{-1} \mathbf{K}_{mn}\mathbf{y}, \\ &k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_{m*}^T \mathbf{K}_{mm}^{-1}\mathbf{k}_{m*} + \\ &\sigma_n^2 \mathbf{k}_{m*}^T \tilde{\mathbf{G}}^{-1}\mathbf{k}_{m*}), \end{aligned}$$

where $\tilde{\mathbf{G}} = \sigma_n^2\mathbf{K}_{mm} + \mathbf{K}_{mn}\mathbf{K}_{nm}$ and $\mathbf{k}_{m*}$ is the vector of kernel evaluations between test input $\mathbf{x}_*$ and the m inducing inputs. The resulting training complexity is $O(m^2 n)$.

In our experiments, the number of inducing points was set empirically by inspecting where the learning curves (in terms of Pearson's correlation gains) flatten, as shown in Figure 1. We used 300 inducing points in experiments with all the settings (see Section 4.3).

4 Experimental Settings

4.1 Data

Our experiments are based on data from two language pairs: English-Spanish (en-es) and English-French (en-fr). The data was collected and made available by WIPO's (World Intellectual Property Organization) Brands and Design Sector. The domain of the data is trademark applications in English, using one or more of the 45 categories of the

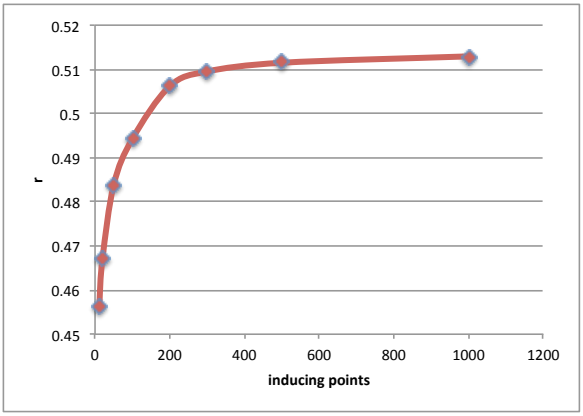


Figure 1: Number of inducing points versus Pearson's correlation

NICE[1] goods and services (e.g. furniture, clothing), and their translations into one of the two languages.

An in-house phrase-based statistical MT system was built by WIPO (Pouliquen et al., 2011), trained on domain-specific data, to translate the English segments. The quality of the translations produced is considered high, with BLEU scores on a 1K-single reference test set reaching 0.71. This is partly attributed to the short length and relative simplicity of the segments in the sub-domains of goods and services. The post-editing was done mostly internally and systematically collected between November 2014 and August 2015. The quality label for each segment is post-editing distance, calculated as the HTER (Snover et al., 2006) between the target segment and its post-edition using the TERCOM tool.[2]

The data was split into 75% for training and 25% for test, with each split maintaining the original data distribution by post-editor. The number of training and test <*source, MT output, post-edited MT, HTER score*> tuples for each of the post-editors (ID) and language pair is given in Table 1. There are 63,763 overlapping English source segments out of 77,656 entries for en-fr and 98,663 entries for en-es. This information is relevant for the crosslingual data experiments, as we discuss in Section 6.

It should be noted that the total number of segments as well as the number of segments per post-editor is significantly higher than those used in previous work. For example, (Cohn and Specia, 2013) used datasets of 6,762 instances (2,254 for each of three translator) and 1,624 instances (299 for each of eight translators), while (Beck et al., 2014) had access to 1000 instances annotated with six emotions.

Lang. Pair	ID	Total	Train	Test
en-es	1	28,423	21,317	7,105
	2	12,904	9,678	3,226
	3	3,939	2,954	984
	4	16,518	12,388	4,129
	5	14,187	10,640	3,546
	6	9,395	7,046	2,348
	7	402	301	100
	8	9,294	6,970	2,323
	9	845	633	211
	10	2,756	2,067	689
	All	98,663	73,997	24,665
en-fr	1	65,280	48,960	16,320
	2	6,336	4,752	1,584
	3	769	576	192
	4	5,271	3,953	1,317
	All	77,656	58,241	19,413

Table 1: Number of en-es and en-fr segments

4.2 Algorithms

For all tasks we used the QuEst framework (Specia et al., 2013) to extract a set of 17 baseline black-box features[3] (Shah et al., 2013) for which we had all the necessary resources for the WIPO domain. These baseline features have shown to perform well in the WMT shared tasks on QE. They include simple counts, e.g. number of tokens in source and target segments, source and target language model probabilities and perplexities, average number of possible translations for source words, number of punctuation marks in source and target segments, among other features reflecting the complexity of the source segment and the fluency of the target segment.

All our models were trained using the GPy[4] toolkit, an open source implementation of GPs written in Python.

4.3 Settings

We built and tested models in the following conditions:

[1] http://www.wipo.int/classifications/nice/en/

[2] http://www.cs.umd.edu/~snover/tercom/

[3] http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox_baseline_17

[4] http://sheffieldml.github.io/GPy/

One language	Setting-1 ind_trn-ind_tst	Setting-2 pol_trn-ind_tst	Setting-3 mtl_trn-ind_tst	Setting-4 ind_trn-pol_tst	Setting-5 pol_trn-pol_tst	Setting-6 mtl_trn-pol_tst
Model	Individual	Pooled	Multitask	Individual	Pooled	Multitask
Test	Individual	Individual	Individual	Pooled	Pooled	Pooled
Crosslingual (cl)	Setting-7 cl_pol_trn-ind_tst	Setting-8 cl_mtl_trn-ind_tst	Setting-9 cl_pol_trn-pol_tst	Setting-10 cl_mtl_trn-pol_tst		
Model	Pooled	Multitask	Pooled	Multitask		
Test	Individual	Individual	Pooled	Pooled		
Non-overlapping (no)	Setting-11 no_cl_pol_trn-pol_tst	Setting-12 no_mtl_trn-pol_tst	Setting-13 no_cl_mtl_trn-pol_tst			
Model	Pooled	Multitask	Multitask			
Test	Pooled	Pooled	Pooled			

Table 2: Various models and test settings in our experiments

- **Setting-1:** Individual models on individual test sets: each model is trained with data from an individual post-editor and tested on the test set for the same individual post-editor.

- **Setting-2:** Pooled model on individual test sets: model trained with data concatenated from all post-editors and tested on test sets of individual post-editors.

- **Setting-3:** Multitask model on individual test sets: multitask models trained with data from all post-editors and tested on test sets of individual post-editors.

- **Setting-4:** Individual models tested on pooled test set: each model is trained with data from an individual post-editor and tested on a test set with data from all post-editors. This setup aims to find out the performance of individual models when the identifier of the post-editor is not known (e.g. in crowdsourcing settings).

- **Setting-5:** Pooled model on pooled test set: model trained with data concatenated from all post-editors and tested on test set of all post-editors.

- **Setting-6:** Multitask model on pooled test set: Multitask model trained with data from all post-editors and tested on test set from all post-editors together.

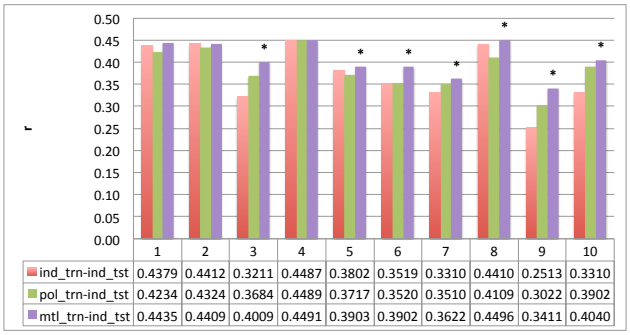
- **Setting-7 to 10:** Similar to setting-2, 3, 5, 6 but with additional crosslingual data where pooled and multitask models are trained with both en-es and en-fr datasets together.

- **Setting-11-13:** Similar to setting-9, 6, 10 respectively, but with non-overlapping crosslingual data only.

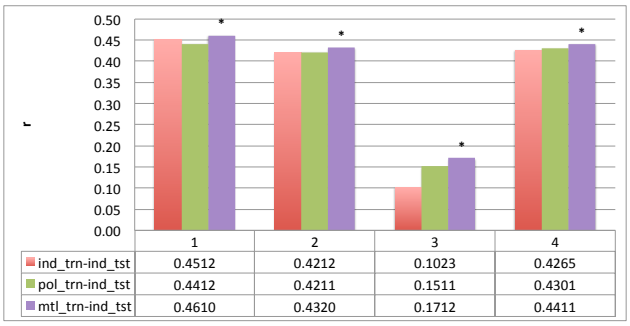
5 Results with Multiple Annotators

We report results in terms of Pearson's correlation between predicted and true quality labels, as was done in the WMT QE shared tasks (Bojar et al., 2015). The multitask learning models consistently led to improvement over pooled models, and over individual models in most cases. We present the comparisons of the models for various settings in the following. The bars marked with * in each comparison are significantly better than all others with $p < 0.01$ according to the Williams significance test (Williams, 1959).

Individual, pooled and multitask models on individual test sets Results for both language pairs are shown in Figure 2. As expected, in cases where a large number of instances is available from an individual post-editor, individual models tested on individual test sets perform better than pooled models. Overall, multitask learning models show improvement over both individual and pooled models, or the same performance in cases where large amounts of data are available for an individual post-editor. For example, in en-es, for post-editors 9 and 3, which have 845 and 3,939 instances in total, respectively, multitask learning models are considerably better. The same goes for post-editor 3 in en-fr, which has only 769 instances. For very few post-editors with a large number of instances (1,2 and 4 in en-es) multitask learning models perform the same as individual or even pooled models. For all other post-editors, multitask models further improve correlation with humans. These results emphasize

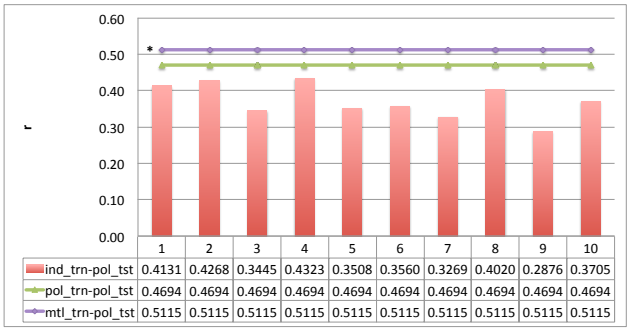


(a) en-es

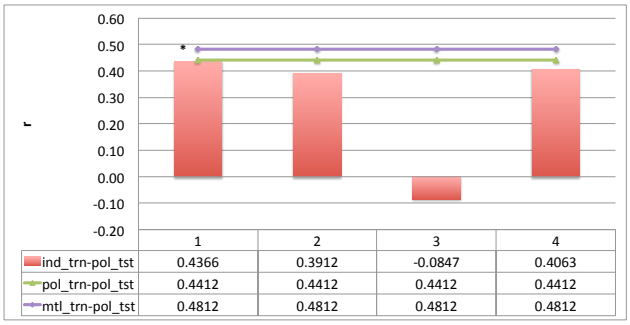


(b) en-fr

Figure 2: Pearson's correlation with various models on individual test sets



(a) en-es



(b) en-fr

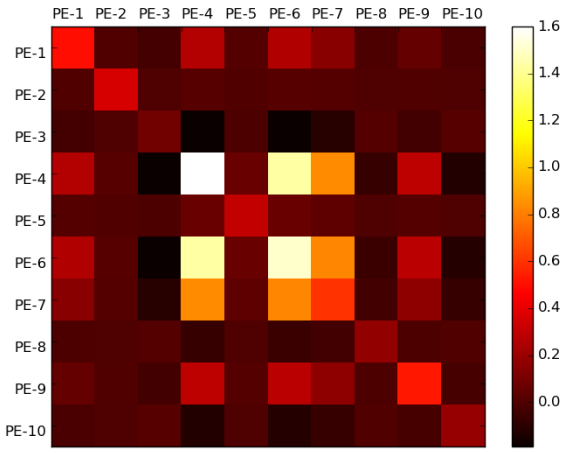
Figure 3: Pearson's correlation with various models on pooled test set

the advantages of multitask learning models, even in cases where the post-editors that will use the models are known in advance (first research question): Clearly, the models for post-editors with fewer instances benefit from the sharing of information from the larger post-editor data sets. As for post-editors with large numbers of instances, in the worst case the performance remains the same.

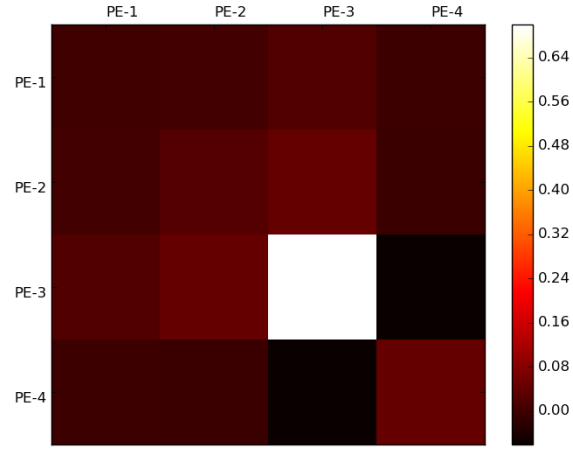
Individual, pooled and multitask models on pooled test set Here we focus on cases where models are built to be used by any post-editor (second research question). Results in Figure 3 show that when test sets for all post-editors are put together, individual models perform distinctively worse than pooled and multitask learning models. Multitask learning models are significantly better than pooled models for both languages (0.511 vs 0.469 for en-es, and 0.481 vs 0.441 for en-fr). In the case of post-editor 3 for en-fr, the correlation is negative for individual models given the very low number of instances for this post-editor, which is not sufficient to build a general enough model that also works for other post-editors.

Relationship among post-editors In order to gain a better insight into the strength of the relationships among various post-editors and thus into the expected benefits from joint modelling, we plot the learned Corregionalisation matrix for all against all post-editors in Figure 4.[5] It can be observed that there exist various degrees of mutual interdependences among post-editors. For instance, in the case of en-es, post-editor 4 shows a strong relationship with post-editors 6 and 7, a relatively weaker relationship with post-editors 1 and 9, and close to non-existing with post-editors 3, 8 and 10. In the case of en-fr, post-editor 3 shows very weak relationship with all other post-editors, especially 4. This might explain the low Pearson's correlation with individual models for post-editor 3 on pooled test sets.

[5]We note that the Corregionalisation matrix cannot be interpreted as a correlation matrix. Rather, it shows the covariance between tasks.



(a) en-es



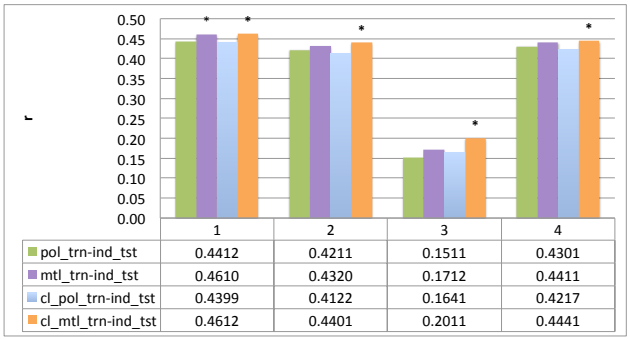
(b) en-fr

Figure 4: Heatmap showing a learned Coregionalisation matrix over all post-editors

6 Results with Multiple Languages

To address the last research question, here we present the results on crosslingual models in comparison to single language pair models. The training models contain data from both en-es and en-fr language pairs in the various settings previously described, where for the multitask settings, tasks can be annotators, languages, or both.

Single versus crosslingual pooled and multitask models on individual test sets Figure 5 shows a performance comparison between single language versus crosslingual models on individual test sets.



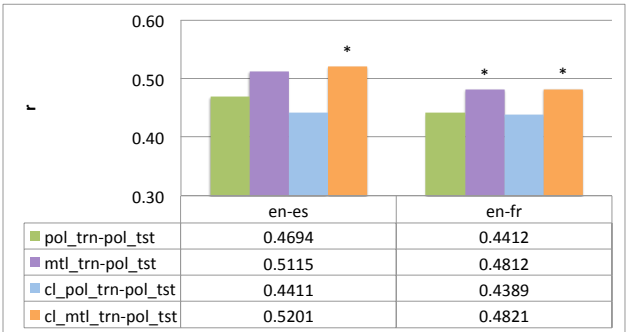
	1	2	3	4
pol_trn-ind_tst	0.4412	0.4211	0.1511	0.4301
mtl_trn-ind_tst	0.4610	0.4320	0.1712	0.4411
cl_pol_trn-ind_tst	0.4399	0.4122	0.1641	0.4217
cl_mtl_trn-ind_tst	0.4612	0.4401	0.2011	0.4441

Figure 5: Pearson's correlation with single versus crosslingual models on individual en-fr test sets

Due to space constraints, we only present results for the en-fr test sets, but those for the en-es test sets follow the same trend. Multitask models lead to further improvements, particularly visible for post-editor 3 (the one with less training data), where the crosslingual multitask learning model reaches 0.201 Pearson's correlation, while the monolingual multitask learning model performs at 0.171. The performance of the pooled models with crosslingual data also improves on this test set over monolingual pooled models, but the overall figures are lower than with multitask learning, showing that the benefit does not only come from adding more data, but from adequate modelling of the additional data. This shows the potential to learn robust prediction models from datasets with multiple languages.

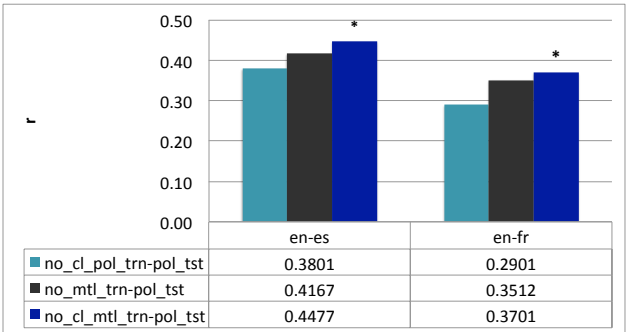
Single versus crosslingual pooled and multitask models on pooled test set Figure 6 compares single language and crosslingual models on the pooled test sets for both languages. A pooled test set with data from different languages presents a more challenging case. Simply building crosslingual pooled models deteriorates the performance over single pooled models, whereas multitask models marginally improve the performance for en-es and keep the performance of the single language models for en-fr. This again shows that multitask learning is an effective technique for robust prediction models over several training and test conditions.

Single versus crosslingual pooled and multitask models on non-overlapping data on pooled test set We posited that the main reason behind the marginal or non-existing improvement of the crosslingual transfer learning shown in Figure 6 is



	en-es	en-fr
pol_trn-pol_tst	0.4694	0.4412
mtl_trn-pol_tst	0.5115	0.4812
cl_pol_trn-pol_tst	0.4411	0.4389
cl_mtl_trn-pol_tst	0.5201	0.4821

Figure 6: Pearson's correlation with single vs crosslingual models: en-es and en-fr pooled test sets



	en-es	en-fr
no_cl_pol_trn-pol_tst	0.3801	0.2901
no_mtl_trn-pol_tst	0.4167	0.3512
no_cl_mtl_trn-pol_tst	0.4477	0.3701

Figure 7: Pearson's correlation with non-overlapping language data: single vs crosslingual models on en-es and en-fr on pooled test sets

the large overlap between the source segments in the datasets for the two language pairs, as mentioned in Section 4: 63,763 instances, which comprise 82% of the en-fr instances, and 65% of the en-es instances. This becomes an issue because nearly half of the quality estimation features are based on the source segments. Therefore, we conducted an experiment with only 41,930 non-overlapping segments in the two languages. This experiment is only possible with pooled test sets, as otherwise too few (if any) instances are left for some post-editors. The results, shown in Figure 7, are much more promising. The Figure compares single language and crosslingual multitask and pooled models on the polled test sets for both languages. It is interesting to note that, while the absolute figures are lower when compared to models trained on all data (Figures 5 and 6), the relative improvements of multitask crosslingual models over multitask single language models are much larger.

7 Conclusions

We investigated multitask learning with GP for QE based on large datasets with multiple annotators and language pairs. The experiments were performed with various settings for training QE models to study the cases where data is available in abundance, versus cases with less data. Our results show that multitask learning leads to improved results in all settings against individual and pooled models. Individual models perform reasonably well in cases where a large amount of training data for individual annotators is available. Yet, by learning from data by multiple annotators, multitask learning models still perform better (in most cases) or at least the same as these models. Testing models on data for individual annotators is a novel experimental setting that we explored in this paper. Another novel finding was the advantage of multitask models in crosslingual settings, where individual models performed poorly and pooled models brought little gain.

Acknowledgments

This work was supported by the QT21 (H2020 No. 645452) and Cracker (H2020 No. 645357). We would like to thank Bruno Pouliquen and Peter Baker for providing the WIPO data, resources and revising the details in Section 4.1.

References

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. 2012. Kernels for Vector-Valued Functions: a Review. *Foundations and Trends in Machine Learning*, pages 1–37.

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

Daniel Beck, Trevor Cohn, and Lucia Specia. 2014. Joint emotion analysis via multi-task gaussian processes. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1798–1803, Doha, Qatar.

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 Sta-

tistical Machine Translation. In *Eigth Workshop on Statistical Machine Translation*, pages 1–44.

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 *Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland.

Ondrej Bojar, Barry Haddow, Matthias Huck, and Philipp Koehn. 2015. Findings of the 2015 workshop on statistical machine translation. In *Tenth Workshop on Statistical Machine Translation*, pages 1–42, Lisboa, Portugal.

Edwin V. Bonilla, Kian Ming A. Chai, and Christopher K. I. Williams. 2008. Multi-task Gaussian Process Prediction. *Advances in Neural Information Processing Systems*.

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 *Seventh Workshop on Statistical Machine Translation*.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28:41–75.

Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. 2010. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11:2901–2934.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *51st Annual Meeting of the Association for Computational Linguistics*, ACL, pages 32–42, Sofia, Bulgaria.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

José G.C. de Souza, Marco Turchi, and Matteo Negri. 2014a. 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.

José G.C. de Souza, Marco Turchi, and Matteo Negri. 2014b. Towards a combination of online and multitask learning for mt quality estimation: a preliminary study. In *Workshop on Interactive and Adaptive Machine Translation*.

José G.C. 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 (Volume 1: Long Papers)*, pages 219–228, Beijing, China.

Jenny Rose Finkel and Christopher D Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610. Association for Computational Linguistics.

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190, Montréal, Canada.

Bruno Pouliquen, Christophe Mazenc, and Aldo Iorio. 2011. Tapta: a user-driven translation system for patent documents based on domain-aware statistical machine translation. In *Proceedings of the 15th conference of the European Association for Machine Translation*, pages 5–12, Leuven, Belgium.

Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian processes for machine learning*, volume 1. MIT Press Cambridge.

Kashif Shah and Lucia Specia. 2014. Quality estimation for translation selection. In *17th Annual Conference of the European Association for Machine Translation*, EAMT, pages 109–116, Dubrovnik, Croatia.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of MT Summit XIV*.

Edward Snelson and Zoubin Ghahramani. 2006. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264.

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 the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

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.

E. J. Williams. 1959. *Regression analysis*. Wiley New York.