TMTprime: A Recommender System for MT and TM Integration

Aswarth Dara[†], **Sandipan Dandapat**[‡,*], **Declan Groves**[†] **and Josef van Genabith**[†]
† Centre for Next Generation Localisation, School of Computing
Dublin City University, Dublin, Ireland
‡ Department of Computer Science and Engineering
IIT-Guwahati, Assam, India
{adara, dgroves, josef}@computing.dcu.ie, sdandapat@iitg.ernet.in

Abstract

TMTprime is a recommender system that facilitates the effective use of both translation memory (TM) and machine translation (MT) technology within industrial language service providers (LSPs) localization workflows. LSPs have long used Translation Memory (TM) technology to assist the translation process. Recent research shows how MT systems can be combined with TMs in Computer Aided Translation (CAT) systems, selecting either TM or MT output based on sophisticated translation quality estimation without access to a reference. However, to date there are no commercially available frameworks for this. TMTprime takes confidence estimation out of the lab and provides a commercially viable platform that allows for the seamless integration of MT with legacy TM systems to provide the most effective (least effort/cost) translation options to human translators, based on the TMTprime confidence score.

1 Introduction

Within the LSP community there is growing interest in the use of MT as a means to increase automation and reduce overall localisation project cost. When high-quality MT output is available, translators see significant productivity gains over translation from scratch, but poor MT quality leads to frustration and wasted time as suggested translations are discarded in favour of providing a translation from scratch. We present a commercially-relevant software platform providing a translation confidence estimation metric and, based on this, a mechanism for effectively integrating MT with TMs in localisation workflows. The confidence metric ensures that only those MT outputs that are guaranteed to require less post-editing effort than the best corresponding TM match are presented to the post-editor (He et al., 2010a). The MT is integrated seamlessly, and established localisation cost estimation models based on TM technologies still apply as upper bounds.

2 Related Work

MT confidence estimation and its relation to existing TM scoring methods, together with how to make the most effective use of both technologies, is an active area of research.

(Specia, 2011) and (Specia et al., 2009, 2010) propose a confidence estimator that relates specifically to the post-editing effort of translators. This research uses regression on both the automatic scores assigned to the MT and scores assigned by post-editors and aims to model post-editors' judgements of the translation quality between *good* and *bad*, or among three levels of post-editing effort.

Our work is an extension of (He et al., 2010a,b,c), and uses outputs and features relevant to the TM and MT systems. We focus on using system external features. This is important for cases where the internals of the MT system are not available, as in the use of MT as a service in a localisation workflow.[1] Furthermore, instead of having to solve a regression problem, our approach is based on solving an easier binary prediction problem (using Support Vector Machines) and can be easily integrated into TMs. (He et al., 2010b) present a MT/TM segment recommender, (He et al., 2010c) a MT/TM n-best list segment re-ranker and (He et al., 2010a) a MT/TM integration method that can use matching sub-segments in MT/TM combination. Importantly,

*Author did this work during his post doctoral research at CNGL.

[1](Specia et al., 2009) note that using glass-box features when available, in addition to black-box features, offer only small gains and also incur significant computational effort.

Proceedings of the NAACL HLT 2013 Demonstration Session, pages 10–13,
Atlanta, Georgia, 10-12 June 2013.

translators can tune the models for precision without retraining the models.

Related research by (Simard and Isabelle., 2009) focuses on combining TM information into an SMT system for improving the performance of the MT when a close match already exists within the TM. (Koehn and Haddow, 2009) presents a post-editing environment using information from the phrase-based SMT system Moses.[2] (Guerberof, 2009) compares the post-editing effort required for TM and MT output, respectively. (Tatsumi, 2009) studies the correlation between automatic evaluation scores and post-editing effort.

3 Translation Recommender

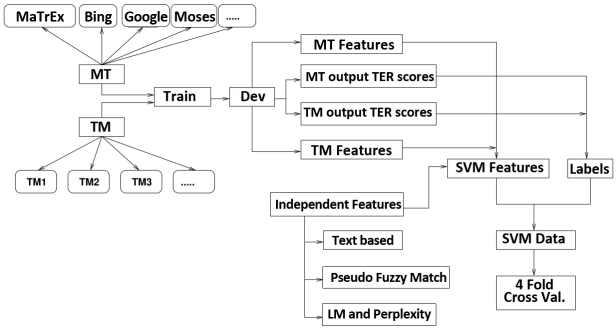

Figure 1: TMTprime Workflow

The workflow of the translation recommender is shown in Figure 1. We train MT systems using a significant portion of the training data and use these models as well as TM outputs to obtain a recommendation development data set. MT systems can be either in-house, e.g. a Moses-based system, or externally available systems, such as Microsoft Bing[3] or Google Translate.[4] For each sentence in the development data set, we have access to the reference as well as to the outputs for each of the MT and TM systems. We then select the best MT (or TM) output as the translation with the lowest TER score with respect to the reference and label the data accordingly. System-independent features for each translation output are fed as input to the SVM classifier (Cortes and Vapnik, 1995). The SVM classifier outputs class labels and the class labels are converted into confidence scores using the techniques given in (Lin et al., 2007). Relying on system independent black-box features has allowed us to build a fully extendable platform that will allow any number of MT systems (or indeed TM systems) to be plugged into the recommender with little effort.

4 Demo Description

Using the Amazon EC2[5] deployment as a back-end, we have developed a front-end GUI for the system (Figure 2). The interface allows the user to select which of the available translation systems (whether they be MT or TM) they wish to use within the recommender system. The user can input their own pre-established estimated cost of post-editing, based on error ranges. Typically the costs for post-editing those translations which have a lower-error rate (i.e. fewer errors) is less than the cost for post-editing translations which have a greater number of errors, as they are of lower quality. The user is requested to upload a file for translation to the system.

Figure 2: TMTprime GUI

Once the user has selected their desired options, the TMTprime platform provides various analysis measures based on its recommendation engine, such as how many segments from the input file are recommended for translation by the various selected translation engines or TMs available. Based on the input costs, it provides a visualisation of overall estimated cost of either using an individual translation system on its own, or using the recommender selecting the best performing system on a segment-by-segment basis. The TMTprime system is an implementation of a segment-based system selector selecting the most appropriate available translation/TM system for a given input. A snapshot of the results produced by TMTprime is given in Figure 3: the pie-chart shows what percentage of segments are recommended from each of the translation systems; the

[2]http://www.statmt.org/moses/
[3]http://www.bing.com/translator
[4]http://translate.google.com/
[5]http://aws.amazon.com/ec2/

bar-graph gives an estimated cost of using a single translation system alone and the estimated cost when using TMTprime's combined recommendation. The estimated cost using TMTprime is lower when compared to using a single MT or TM system alone (in the worst case, it will be the same as the best-performing single translation engine or TM system). This estimated cost includes both the cost for translation (currently uniform cost for each translation system) and the cost required for post-editing. For example, if the MT is an in-house system the cost of translation will be (close to) zero whereas there is potentially an additional base cost for using an external MT engine. Finally, the interface provides statistics related to various confidence levels for different translation outputs across the various translation and TM systems.

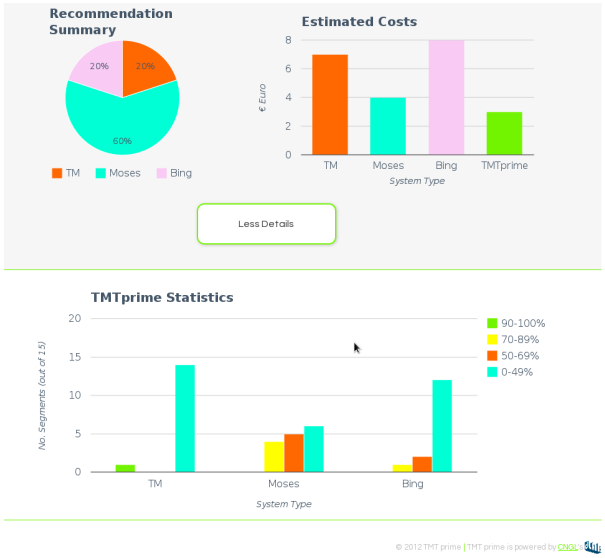

Figure 3: Results shown by TMTprime system

5 Experiments and Results

Evaluation targets two objectives and is described below.

5.1 Correlation with Automatic Metrics

TER and METEOR are widely-used automatic metrics (Snover et al., 2006; Denkowski and Lavie, 2011) that calculate the quality of translation output by comparing it against a human translation, known as the reference translation. Our data sets for the experiment consist of English-French translation memories from the IT domain. In all instances MT was carried out for English-French translations. As we have access to the reference target language translations for our test set, we are able to calculate the TER and METEOR scores for the three translation outputs (here TM, MaTrEx (Dandapat et al., 2010) and Microsoft Bing). For each sentence in the test set, TMTprime recommends a particular translation output with a certain estimated confidence level without access to a reference. We measure Pearson's correlation coefficient (Hollander and Wolfe, 1999) between the recommendation scores, TER scores and METEOR scores (for all system outputs) in order to determine how well the TMTprime prediction score correlates with the widely used automatic evaluation metrics. Results of these experiments are provided in Table 1 which shows there is a negative correlation between TMTprime scores and TER scores. This shows that both TMTprime scores and TER scores are moving in opposite directions, supporting the claim that the higher the recommendation scores, the lower the TER scores. As TER is an error score, the lower the TER score, the higher the quality of the machine translation output compared to its reference. On the other hand, TMTprime scores are positively correlated with METEOR scores which supports the claim that the higher the recommendation scores, the higher the METEOR scores.

Pearson's r	**TER**	**METEOR**
TMTprime	-0.402	0.447

Table 1: Correlation with automatic metrics

The evaluation has been performed on a test data set of 2,500 sentences. Both the correlations are significant at the ($p<0.01$) level.

5.2 Correlation with Post-Editing time

This is the most important and crucial metric for the evaluation. For this experiment we made use of post-editing data captured during a real-world translation task, for English-French in the IT domain.

Pearson's r	**TER**	**METEOR**	**PE Time**
TMTprime	-0.122	0.129	-0.132

Table 2: Correlation with Post-Editing times

For testing, we collect the post-editing times for MT outputs from two different translators using a commercial computer-aided translation (CAT tool) in a real-world production scenario. The data set consists of 1113 samples and is different from the one used in the correlation with automatic metrics.

Post-editing times provide a real measure of the amount of post-editing effort required to perfect the output of the MT system. For this experiment, we took the output of the MT system used in the task together with the post-editing times and measured the Pearsons correlation coefficient between the TMTprime recommendation scores and the post-editing (PE) times (only for MT output from a single system since this data set does contain PE times for other translation outputs). In addition, we also repeated the previous experiment setup for finding the correlation between the TMTprime scores and the automatically-produced TER, METEOR scores for this data set. The results are given in Table 2.

The results show that the confidence scores do correlate with automatic evaluation metrics and post-editing times. Although the correlations do not seem as strong as before, the results are statistically significant ($p<0.01$).

6 Conclusions and Future Work

We present a commercially viable translation recommender system which selects the best output from multiple TM/MT outputs. We have shown that our confidence score correlates with automatic metrics and post-editing times. For future work, we are looking into extending and evaluating the system for different language pairs and data sets.

Acknowledgments

This work is supported by Science Foundation Ireland (Grants SFI11-TIDA-B2040 and 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. We would also like to thank Symantec, Autodesk and Welocalize for their support and provision of data sets used in our experiments.

References

Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. In *Machine Learning*. pages 273–297.

Dandapat, Sandipan, Mikel L. Forcada, Declan Groves, Sergio Penkale, John Tinsley, and Andy Way. 2010. OpenMaTrEx: A free/open-source marker-driven example-based machine translation system. In *Proceedings of the 7th international conference on Advances in natural language processing*. Springer-Verlag, Berlin, Heidelberg, IceTAL'10, pages 121–126.

Denkowski, Michael and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*. Edinburgh, UK.

Guerberof, Ana. 2009. Productivity and quality in mt postediting. In *Proceedings of Machine Translation Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators*. Ottawa, Canada.

He, Yifan, Yanjun Ma, J Roturier, Andy Way, and Josef van Genabith. 2010a. Improving the post-editing experience using translation recommendation: A user study. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*. Denver, Colorado, AMTA 2010, pages 247–256.

He, Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010b. Bridging smt and tm with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, ACL 2010, pages 622–630.

He, Yifan, Yanjun Ma, Andy Way, and Josef van Genabith. 2010c. Integrating n-best smt outputs into a tm system. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, Beijing, China, COLING 2010, pages 374–382.

Hollander, Myles and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. John Wiley and Sons.

Koehn, Philip and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of MT Summit XII*. Ottawa, Canada, pages 73–80.

Lin, Hsuan-Tien, Chih-Jen Lin, and Ruby C. Weng. 2007. A note on platt's probabilistic outputs for support vector machines. *Machine Learning* 68(3):267–276.

Simard, Michael and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. In *Proceedings of Machine Translation Summit XII*. Ottawa, Canada, pages 120–127.

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

Specia, Lucia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*. Leuven, Belgium, EAMT 2011, pages 73–80.

Specia, Lucia, Nicola Cancedda, and Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *Proceedings of LREC 2010*. Valletta, Malta.

Specia, Lucia, Marco Turqui, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009. Improving the confidence of machine translation quality estimates. In *Proceedings of Machine Translation Summit XII*. Ottawa, Canada, pages 136–143.

Tatsumi, Midori. 2009. Correlation between automatic evaluation scores, post-editing speed and some other factors. In *Proceedings of Machine Translation Summit XII*. Ottawa, Canada, pages 332–339.