Combining Machine Translation Systems with Quality Estimation

László János Laki[1,3] and Zijian Győző Yang[2]

[1] MTA-PPKE Hungarian Language Technology Research Group
[2] Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
Práter str. 50/A, 1083 Budapest, Hungary
[3] MorphoLogic Lokalizáció Kft.
Logodi str. 54, 1012 Budapest, Hungary
{laki.laszlo, yang.zijian.gyozo}@itk.ppke.hu

Abstract. Improving the quality of Machine Translation (MT) systems is an important task not only for researchers but it is a substantial need for translating companies to create translations in a quicker and cheaper way. Combining the outputs of more than one machine translation systems is a common technique to get better translation quality because the strengths of the different systems could be utilized. The main question is to find the best method for the combination. In this paper, we used the Quality Estimation (QE) technique to combine a phrase-based and a hierarchical-based machine translation systems. The composite system was tested on several language combinations. The QE module was used to compare the outputs of the different MT systems and gave the best one as the result translation of the composite system. The composite system gained better translation quality than the separated systems.

1 Introduction

In the past few years Machine Translation (MT) systems have undergone significant changes. The goal of the researchers was to create better and better translations, which lead them to implement numerous MT systems based on the actual technological brands (e.g. rule-based MT, statistical MT, syntactical MT, neural MT). These methods have different advantages, therefore these systems could be used for different problems with high efficiency. For example, the rule-based MT system is more efficient when translating between inflected languages, because it is able to generate inflected word forms based on language specific grammar rules. However it is not able to translate unknown words effectively, which is the strength of SMT systems. Combining different kinds of MT systems can join their advantages and reduce the deficiency of the systems. The combined result can achieve better quality than the original MT systems.

Recently, there have been several experiments in combining outputs from different MT systems [1, 10, 13, 16]. These combinations can be realized in many ways. One of these methods is to build a word-level confusion network from hypotheses translations [12], others work on sentence-level [10]. There are differences in the alignment methods used for the generation of the confusion network [6, 15, 17, 18, 20].

The novelty of our system is that we used sentence level quality estimation (QE) calculation for the phrase-based (PB) and hierarchical-based (HB) MT system outputs to choose the best translation. The QE (see Figure 1) estimates the quality of the MT translated segment without reference translation. It is based on a statistical model trained by regression analysis. Quality indicators are extracted based on the source segment, the machine translated segment and inner parameters of the MT system. The QE model is trained based on these indicators and on human or automatic evaluation scores. After all, using the trained statistical model, we can predict the quality of the new unseen segments. In our research we translated the source segments with two different MT systems, then using a QE model we chose the better quality translation as the final MT output.

The structure of this paper is as follows: First, we discuss the related work. Then, we shortly introduce the quality estimation approach. After this, we explain our methods and experiments in detail. Finally, our results and conclusions are described.

2 Related Work

The most common combination method is using confusion network decoding method to combine and choose the best translation outputs from multiple MT systems. A word-level alignment method is to select a monolingual reference against a hypothesis. Then, using this alignment, a confusion network is built from the hypotheses [13, 19]. For this task, Rosti et al. [18] and Heafield et al. [6] used the TER [20] algorithm, Okita et al. [15] used the BLEU [17] method. Rosti et al. used an incremental word-based alignment method to build a confusion network. The alignment is based on the TER algorithm. This incremental alignment uses a pair-wise hypothesis. All alternative translation hypotheses are aligned against a ”skeleton” hypothesis independently. Then, using the incremental alignment decoding, confusion networks are built from the union of hypotheses alignments. Heafield et al. [6] improved the confusion network decoding with phrase-level alignment. Huang and Papineni [7] created a hierarchical system combination strategy. This approach can combine MT hypotheses on word, phrase and sentence levels.

In our research, we used the QE approach to combine the MT outputs.

3 Quality Estimation

In the QE [21] task (see Figure 1), we extract different kinds of quality indicators from the source and translated sentences without using reference translations. Following the research of Specia et al. [21], we can separate different feature categories. The first class contains the complexity features (e.g. number of tokens in the source segment), which are extracted from the source sentences. To the second class we extract fluency features (e.g. percentage of verbs in the target sentences) from the translated sentences. The third category includes the adequacy features (e.g. ratio of percentage of nouns in the source and target) extracted from the comparison of the source and the translated sentence. Finally, we can also extract features from the decoder of the MT system, namely confidence features (e.g. features and global score of the SMT system).

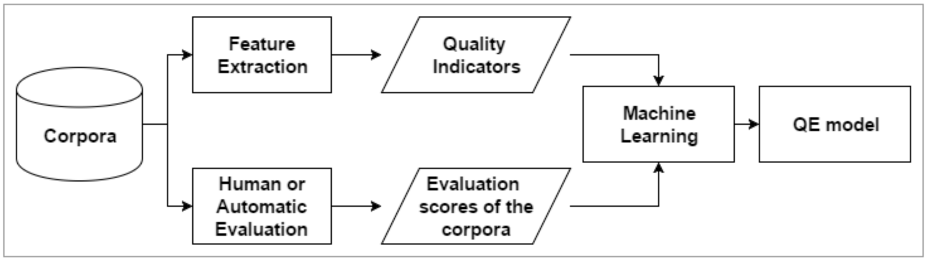

Fig. 1. Quality Estimation model

From another perspective the features could be divided into two main groups: "black-box" and "glass-box" features. The features extracted from the inner parameters of the MT system are called "glass-box" features. Definition of these features requires access to the MT system, but in most cases the internal accessibility of the systems is not allowed. The features which do not depend on the MT inner parameters are the "black-box" features. These features make decisions based on the source and the hypothesis translation only. Since in our experiments we have translations from different MT systems, we use only the "black-box" features.

Thanks to the regression analysis, the QE system is able to predict the quality of the MT output compared to any measure score using the extracted quality indicators and features. Ideally, this measure would be human evaluation, but it is expensive and time-consuming. That is why standard automatic metrics (e.g. BLEU [17], OrthoBLEU [5], TER [20], etc.) are used to determine the quality of a translation. One of the advantages of the QE model is that it does not require reference translations for quality prediction, therefore it is an applicable and fully automatic solution to combine different kind of MT systems.

4 Introducing the used MT systems

Thanks to the QE technique, it is possible to create any number and any types of multiple MT systems. In this paper we used phrase-based (PB) and hierarchical-based (HB) statistical machine translation systems.

PBMT [11] systems rely on statistical observations derived from those phrase alignments which are automatically extracted from parallel bilingual corpora. The main reason to use SMT is its language independent behavior, which can be used successfully in the case of language pairs with similar syntactic structure and word order. PBMT is a solution to handle local reordering, but it has difficulties with long distant ones. HBMT [3] tries to solve this reordering issue.

HBMT is the extension of the PBMT. While PBMT uses a phrase-to-phrase stack decoder, HBMT uses context free grammar based chart decoder. This technique helps HBMT to learn reordering patterns in a partially lexicalized form. For example the English-French negation is stored as *don't* $X \rightarrow$ *ne* X *pas*, where X can be replaced by any verb phrase.

If we compare the PB an HB systems, we can observe that their performance highly depends on the language pairs and the domain of the corpus. This explains why a HB system is not able to outperform the PB system in all cases (shown in Table 2). Consequently more robust MT systems can be created by choosing the better translation from the outputs of these two MT systems.

5 Methods and Experiments

In this section we will describe our experimental setups. First of all we will show the settings of the translation systems and after that the settings of the quality estimation system. Finally the system combination will be presented.

5.1 Dataset

We performed experiments on four language pairs, where English was the source language and the target languages were Hungarian, German, Italian and Japanese. These language combinations gave us a wide overview of the performance of the system, since the structure of these languages is very different. The wide overview was also supported by the used corpus, which was built from four domains (car industry documentation, European Parliament documentations, IT and industrial product documentation). The IT or the industrial documents contained mostly short segments, while the segments in law texts usually included more than 20 words. The biggest parts of the corpora were given by a translation company, therefore this part is not open-source. In order to make our experiments reproducible, one of our corpora was an open-source one: The Acquis Communautaire multilingual parallel corpus[4] used in the case of English-German translation. The exact size of the resources are shown in Table 1. The segments in each text are unique without repetitions. Also there are no overlapped segments between the train and the test sets.

5.2 Machine Translation system setup

First of all, preprocessing was made on the training set, which included tokenization and truecaseing. The word alignment was created with GIZA++ [14].

[4] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

Table 1. Corpora used in the experiment

Language pair	Domain	# of MT training segments	# of QE training segments	# of QE test segments
English - Hungarian (en-hu)	car industry	240,000	6,000	1,500
English - German (en-de)	law	1,000,000	5,250	1,300
English - Italian (en-it)	product description	800,000	3,143	785
English - Japanese (en-ja)	IT	1,000,000	3,169	790

Both phrase-based and hierarchical machine translation systems were realized by Moses [9]. This system was used for building 5-gram translation models as well. The 3-gram language models were built by the IRSTLM [4] tool. Our translation system could handle XML markups correctly, thanks to the *m4loc* architecture [8]

5.3 Quality Estimation system setup

In our research, we used automatic evaluation metrics for building QE models, i.e. BLEU [17], OrthoBLEU [5] and OrthoTER scores. It means that our QE systems will predict a float number between 0 and 1 based on these metrics. The OrthoBLEU and OrthoTER methods work on character level, therefore these perform better than BLEU in the case of agglutinating and compounding languages like Hungarian. If the translation fails only at inflections, the BLEU gives a low-score, even if the stem is translated correctly. In such cases OrthoBLEU scores are more accurate.

As it was presented in Section 3, feature extraction is needed to build the QE model. For this task we applied the QuEst [21] system. We used only "black box" and language independent features to create the proper QE models for both MT systems for all four languages. In our research 67 features were applied, which were developed by Specia et. al. [21]. These 67 features contain adequacy and fluency features (number of tokens in source and target segments, percentage of source 1–3-grams observed in different frequency quartiles of the source side of the MT training corpus, average number of translations per source word in the segment as given by IBM-1 model [2], etc.).

For QE model training, we tried several regression methods, for example support vector regression, decision trees and rules, Gaussian process etc. The Gaussian process (GP) with RBF Kernel gained the highest performance, thus in the results section we show only the GP scores.

For Hungarian, an optimization task (referred to as *en-hu-opt*) has been applied. According to the research of Yang et. al [23, 22] an additional 60 features were added, from which 53 were developed by the authors of this paper. These features were language specific features (ratio of percentage of verbs and nouns in the target, percentage of nouns in the target, etc.), n-gram features (perplexity of the target, log probability of the target, etc.), error features (percentage of unknown words in the target, percentage of XML tags in the target, etc.) and semantic features (WordNet features, dictionary features, etc.). In the evaluation section we will compare the results of the basic and the optimized systems.

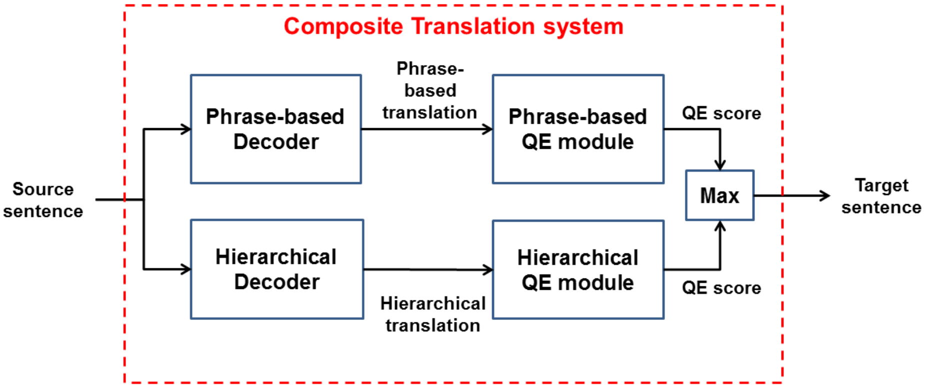

Fig. 2. Composite decoder architecture

5.4 Composite Translation System

The composite translation system was built from the PBMT and HBMT systems. The architecture of the system is shown in Figure 2. First of all, the input segment translation is processed by the PBMT and the HBMT modules, then the appropriate QE modules predict an evaluation score for these translations. Finally, the system chooses the better translation based on the estimated rate, which will be the final output of the composite system.

6 Results and Evaluation

In our experiments, the QE train and test sets were translated with both MT systems. Based on the source and the translated sentences, the QE features have been extracted, then the GP regression was trained on one of the automatic evaluation metrics. Finally, the QE scores of the test sets were predicted. These steps were performed on the four language pairs and the three evaluation metrics. In Table 2, we can see the BLEU, OrthoBLEU and OrthoTER scores of the PBMT and the HBMT systems separately and the scores of the composite MT (Co MT). One more row is shown, namely the Max MT, which means the theoretical maximum score for Co MT translation, if the QE model would be perfect. Co MT system outperforms PBMT and HBMT systems in every cases.

During the deeper evaluation of the composite system, we took a closer look at the performance of the translation selection module. The precision, the recall and the F-measure were calculated both for the PB and for the HB selection cases

Table 2. Evaluation scores of the combined MT systems

		en-hu	en-hu-opt	en-de	en-it	en-ja
BLEU mean score ↑	PB MT	0.5156	0.5156	0.6288	0.7513	0.5945
	HB MT	0.6157	0.6157	0.4808	0.6998	0.6044
	Co MT	**0.6360**	**0.6375**	**0.6302**	**0.7525**	**0.6057**
	max MT	0.6702	0.6702	0.6475	0.7660	0.6458
OrthoBLEU mean score ↑	PB MT	0.7381	0.7381	0.6757	0.8202	0.5361
	HB MT	0.7679	0.7679	0.6221	0.7993	0.5536
	Co MT	**0.7795**	**0.7788**	**0.6788**	**0.8246**	**0.5553**
	max MT	0.8023	0.8023	0.6979	0.8374	0.5832
OrthoTER mean score ↓	PB MT	0.2903	0.2903	0.3574	0.1669	0.4281
	HB MT	0.2193	0.2193	0.4170	0.1995	0.4075
	Co MT	**0.2085**	**0.2108**	**0.3540**	**0.1662**	**0.4055**
	max MT	0.1848	0.1848	0.3349	0.1542	0.3769

as well. With these statistics, the performance of the quality predication could be measured. The last row contains the system level accuracy of the selection method, which means how many times the system chose the right one of the compared translations. For the calculation of F-measure, we counted as a positive predication when the PB and the HB translations were identical. This is the reason why F-measure could be higher than accuracy in the case of both systems.

These results are shown in Table 3. In most cases, our QE models are able to estimate with high accuracy. It is interesting to see that the precision of the selector metric is above 80% in the case of word-level metric and it is above ∼ 72% in the case of character level measures. From these numbers we could see that the problem is with the recall of the MT system which has the lower quality. In most cases recall is near 65%, which could be an answer for the decrease of the accuracy of the Co MT system.

In the case of the comparison of the basic Hungarian model to the optimized Hungarian model, we used the statistical correlation, the MAE (Mean absolute error) and the RMSE (Root mean-squared error) evaluation metrics. The correlation ranges are from -1 to +1; the correlation is better if it is closer to the edge of the range. In the case of MAE and RMSE metrics, closer values to 0 mean a better QE model.

In Table 2 we can see that in the case of the BLEU metric ∼ 2% higher correlation was reached with the optimized feature set. It could also be noticed, that in Table 4 the optimized Hungarian BLEU model could make higher accuracy prediction than the basic Hungarian BLEU model. The features we used for optimization were word-based features, hence only the optimized BLEU model could gain higher accuracy.

7 Conclusion

We created a composite machine translation system, which combines the output of multiple machine translation systems based on sentence-level quality estima-

Table 3. Evaluation of the performance of the composite system

			en-hu	en-hu-opt	en-de	en-it	en-ja
BLEU	Precision	PB	88.224%	85.662%	85.229%	93.808%	90.846%
		HB	81.707%	82.979%	96.790%	97.513%	86.483%
	Recall	PB	64.809%	68.328%	98.810%	98.072%	84.976%
		HB	94.783%	93.103%	67.703%	92.114%	91.821%
	F-measure	PB	74.725%	76.020%	91.518%	95.892%	87.813%
		HB	87.761%	87.750%	79.675%	94.737%	89.072%
	System accuracy		80.067%	80.400%	84.615%	92.229%	81.519%
OrthoBLEU	Precision	PB	81.261%	78.547%	71.736%	90.997%	92.647%
		HB	79.834%	80.394%	97.324%	94.679%	85.894%
	Recall	PB	64.986%	67.003%	98.794%	95.773%	83.306%
		HB	90.244%	88.086%	52.980%	88.812%	93.893%
	F-measure	PB	72.218%	72.317%	83.118%	93.324%	87.728%
		HB	84.720%	84.064%	68.611%	91.652%	89.716%
	System accuracy		76.867%	76.267%	71.846%	88.025%	82.152%
OrthoTER	Precison	PB	85.714%	82.305%	73.132%	91.081%	94.559%
		HB	80.168%	80.833%	97.066%	95.644%	86.915%
	Recall	PB	59.627%	62.112%	98.712%	96.700%	84.140%
		HB	94.260%	92.287%	54.014%	88.441%	95.606%
	F-measure	PB	70.330%	70.796%	84.018%	93.807%	89.046%
		HB	86.645%	86.181%	69.406%	91.902%	91.053%
	System accuracy		78.400%	78.000%	73.077%	88.662%	84.304%

Table 4. Optimized BLEU Hungarian model

			en-hu	en-hu-opt
BLEU	Correlation	PB	0.6667	**0.6884**
		HB	0.5926	**0.6199**
	MAE	PB	0.1809	**0.1730**
		HB	0.1953	**0.1888**
	RMSE	PB	0.2266	**0.2196**
		HB	0.2402	**0.2341**

tion. In our experiments, phrase-based and hierarchical-based MT systems were combined, but with this technique any kind and number of systems could be combined. Quality estimation method with "black-box" features was used for the combination. The composite system was tested on four different language pairs. Results showed that our Co MT system gained better final translation quality compared to PBMT and HBMT systems in any experiments. In the case of English-Hungarian language pairs, some language dependent QE features were integrated to the Hungarian QE model, which led us to better prediction.

8 Acknowledgement

We would like to thank MorphoLogic Lokalizáció Kft. for the data sets which were used in our research.

References

1. Bangalore, S., Bordel, G., Riccardi, G.: Computing consensus translation from multiple machine translation systems. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 350–354. Madonna di Campiglio, Italy (2001)
2. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Comput. Linguist. 19(2), 263–311 (Jun 1993)
3. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 263–270. ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA (2005)
4. Federico, M., Bertoldi, N., Cettolo, M.: IRSTLM: an open source toolkit for handling large scale language models. In: INTERSPEECH. pp. 1618–1621 (2008)
5. FTSK: Orthobleu mt evalution based on orthographic similarities @ONLINE (May 2014)
6. Heafield, K., Hanneman, G., Lavie, A.: Machine translation system combination with flexible word ordering. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. pp. 56–60. Association for Computational Linguistics, Athens, Greece (March 2009)
7. Huang, F., Papineni, K.: Hierarchical system combination for machine translation. In: EMNLP-CoNLL. pp. 277–286 (2007)
8. Hudk, T., Ruopp, A.: The integration of Moses into localization industry. In: Proceedings of the 15th International Conference of the European Association for Machine Translation. Leuven, Belgium (2011)
9. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL. pp. 177–180 (2007)
10. Kumar, S., Byrne, W.: Minimum bayes-risk decoding for statistical machine translation. In: Susan Dumais, D.M., Roukos, S. (eds.) HLT-NAACL 2004: Main Proceedings. pp. 169–176. Association for Computational Linguistics, Boston, Massachusetts, USA (2004)

11. Lopez, A.: Statistical machine translation. ACM Comput. Surv. 40(3), 8:1–8:49 (Aug 2008)
12. Mangu, L., Brill, E., Stolcke, A.: Finding consensus among words: lattice-based word error minimization. In: Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999. pp. 495–498 (1999)
13. Matusov, E., Ueffing, N., Ney, H.: Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In: McCarthy, D., Wintner, S. (eds.) EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy. The Association for Computer Linguistics (2006)
14. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
15. Okita, T., van Genabith, J.: Minimum Bayes Risk Decoding with Enlarged Hypothesis Space in System Combination, pp. 40–51. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
16. Okita, T., Rubino, R., Genabith, J.v.: Sentence-level quality estimation for mt system combination. In: Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT. pp. 55–64. The COLING 2012 Organizing Committee, Mumbai, India (December 2012)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
18. Rosti, A.V., Zhang, B., Matsoukas, S., Schwartz, R.: Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In: Proceedings of the Third Workshop on Statistical Machine Translation. pp. 183–186. Association for Computational Linguistics, Columbus, Ohio (June 2008)
19. Sim, K.C., Byrne, W.J., Gales, M.J.F., Sahbi, H., Woodland, P.C.: Consensus network decoding for statistical machine translation system combination. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. vol. 4, pp. IV–105–IV–108 (April 2007)
20. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: In Proceedings of Association for Machine Translation in the Americas. pp. 223–231 (2006)
21. Specia, L., Shah, K., de Souza, J.G., Cohn, T.: Quest - a translation quality estimation framework. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 79–84. Sofia, Bulgaria (2013)
22. Yang, Z.G., Laki, L.J.: Minőségbecslő rendszer egynyelvű természetes nyelvi elemzőhöz. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 37–49. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged, Hungary (2017)
23. Yang, Z.G., Laki, L.J., Siklsi, B.: Quality estimation for English-Hungarian with optimized semantic features. In: Computational Linguistics and Intelligent Text Processing. Konya, Turkey (2016)