HuQ: An English-Hungarian Corpus for Quality Estimation

Zijian Győző Yang[*], **László János Laki**[†], **Borbála Siklósi**[†]
[*]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
[†]MTA-PPKE Hungarian Language Technology Research Group
Práter str. 50/A, 1083 Budapest, Hungary
{yang.zijian.gyozo, laki.laszlo, siklosi.borbala}@itk.ppke.hu

Abstract

Quality estimation for machine translations is an important task. The standard automatic evaluation methods that use reference translations cannot perform the evaluation task well enough. These methods are low correlated with human evaluations in the case of English-Hungarian translation. Quality estimation is a new approach to solve this problem. This method is a prediction task by estimating the quality of translations for which features are extracted from only the source and translated sentences. The quality estimation was not implemented for Hungarian before, thus there is no training corpus. In this study, we created a dataset to build quality estimation models for English-Hungarian. We also did experiments to optimize the quality estimation system to Hungarian. In the optimization task we did research in the field of feature engineering and feature selection. We created optimized feature sets, which produced better results than the baseline feature set.

Keywords: quality estimation, machine translation, HuQ

1. Introduction

The measurement of quality of translation output has become necessary. Especially in the field of machine translation (MT). A reliable quality score for MT could save a lot of time and money for translators, companies, researchers and ordinary users. Knowing the quality of machine translated segments can accelerate the translators' work, or can help human annotators in their post-edit tasks, or can filter out and inform about unreliable translations. Last but not least, quality indicators can help MT systems to combine the translations to produce better output. There are two kinds of evaluation methods for MT. The first type uses reference translations, i.e. it compares machine translated sentences to human translated reference sentences, and measures the similarities or differences between them. To know the quality of MT, after an automatic translation, we also have to create a human translated sentence (for the sentences of the test set) to compare it to the machine translated output. Creating human translations is expensive and time-consuming, thus these methods, such as BLEU (Papineni et al., 2002) and other methods based on BLEU, TER (Snover et al., 2006), HTER (Snover et al., 2006) etc., cannot evaluate in run-time, and the correlation between the results of these methods and that of human evaluation is very low in the case of translations from English to Hungarian. A completely new approach is needed to solve these problems, i.e. a method which can predict translation quality in real-time and does not need reference translations.

The other type of evaluation methods is called Quality Estimation (QE). This is a supervised approach that does not use reference translations. This method addresses the problem by evaluating the quality of machine translated segments as a prediction task. Using QE we can save considerable time and money for translators, human annotators, researchers, companies and ordinary users.

In this study, we use the QuEst framework (Specia et al., 2013), developed by Specia et al., to train and apply QE models for Hungarian, which to our knowledge has not been done before. Hence, first, we needed to create a QE corpus for Hungarian. Then, using this corpus we built different kinds of optimized English-Hungarian QE models. For optimizing we developed new semantic features using WordNet and word embedding models.

Hungarian is an agglutinating and compounding language. There are significant differences between English and Hungarian, regarding their morphology, syntax and word order or number. Furthermore, the free order of grammatical constituents, and different word orders in noun phrases (NPs) and prepositional phrases (PPs) are also characteristics of Hungarian. Thus, features used in a QE task for English-Spanish or English-German, which produced good results, perform much worse for English-Hungarian. Hence, if we would like to use linguistic features in QuEst, we need to integrate the available Hungarian linguistic tools into it.

The structure of this paper is as follows: First we will shortly introduce the QE approach. Then, we will present the corpus we created for English-Hungarian QE. Finally, our experiments, optimizations and results in the task of QE are described.

2. Related Work

In the last couple of years there have been several WMT workshops with quality estimation shared tasks,[1] which provided datasets for QE researches. The datasets are evaluated with HTER, METEOR, ranking or post-edit effort scores. But, unfortunately, there is no dataset for Hungarian. In this research we created a QE dataset for English-Hungarian. For human judgement we used a general scoring scale.

QE is a prediction task, where different quality indicators are extracted from the source and the machine translated segments. The QE model is built with machine learning algorithms based on these quality indicators. Then the QE

[1] http://www.statmt.org/wmt15/quality-estimation-task.html

model is used to predict the quality of unseen translations. The aim is that the scores, predicted with the QE model highly correlate with human judgments, thus the QE model is trained on human evaluations.

In the recent years, in the field of QE, research has focused on feature selection (Biçici, 2013) using a variety of machine learning algorithms and feature engineering (Camargo de Souza et al., 2013). In feature selection task, Beck et al. tried more than 160 features in an experiment for English-Spanish to predict HTER (Beck et al., 2013). Recently, research (Bojar et al., 2015) in QE has focused on providing larger datasets; feature selection using a variety of machine learning algorithms and feature engineering for word-level, sentence-level and document level QE; exploring differences between sentence-level and document-level prediction; and analyzing training data size and quality. In our research we did experiments for Hungarian QE in providing a dataset, word-level feature engineering and feature selection.

3. Quality Estimation

In the QE task, we extract different kinds of features as quality indicators from the source and translated sentences. Following the research of Specia et al., we can separate the features in different kinds of category (Specia et al., 2013). From the source sentences, complexity features can be extracted (e.g. number of tokens in the source segment). From the translated sentences, we extract fluency features (e.g. percentage of verbs in the target sentences). From the comparison between the source and the translated sentences, adequacy features are extracted (e.g. ratio of percentage of nouns in the source and target). We can also extract indicators from the MT system, these are the confidence features (e.g. features and global score of the SMT system). From another point of view, we can also divide the features into two main categories: "black-box" features (independent from the MT system) and "glass-box" features (MT system-dependent). Since in our experiments we have translations from different MT systems, we did use only the "black-box" features. After feature extraction, using these quality indicators, we can build QE models with machine learning methods. The aim is that the predictions of the QE models are highly correlated with human evaluations. Thus, the extracted quality indicators need to be trained on human judgments.

4. HuQ Corpus

To build the English-Hungarian QE system, we needed a training corpus. In our experiments, we created a corpus called Hungarian QE (HuQ). The HuQ corpus contains 1500 English-Hungarian sentence pairs. To build the HuQ corpus, we used 300 English sentences of mixed topics from the Hunglish corpus (Halácsy et al., 2005). We translated these 300 sentences into Hungarian with different MT systems. After the translation, to create human judgements, we evaluated these translated segments with human annotators. For creating human scores, we developed a website[2] with a form for human annotators to evaluate the translations. In this website we can see an English source sentence and its Hungarian translation, originating from one of the translation sources. However, the evaluates were not aware of the origin of the translation. The annotators could give quality scores from 1 to 5, from two points of view (Koehn, 2010): adequacy and fluency (see Table 2). We added a 0 score (*I do not understand the English sentence*) to filter out wrong evaluations. All the 1500 sentences were evaluated by 3 human annotators: L, M and T. All the annotators were native Hungarian speakers who have minimum B2 level English language skill. The 3 annotators have different evaluation point of view:

- L: linguist,
- M: MT specialist,
- T: language technology expert.

To follow and control the annotators effectively, or to discuss the annotation aspects with the annotators personally, to avoid misunderstandings, we did not use crowd sourcing for the evaluation. For the consistency of measurement of calculating agreement, the 3 annotators evaluated a set of 50 translations in a personal meeting. These translations are not included in the training set.

There are 3 topics in the HuQ corpus: subtitles, literature and law. The subtitles are simple daily used sentences containing a high ratio of slang words. The language of literature has more complex grammatical constructions with many rare words used. The segments from law are official texts with complex grammar.

We used 5 different MT system to translate each of the 300 sentences:

1. Human translation from the Hunglish corpus,
2. MetaMorpho (Novák et al., 2008) rule based MT system,
3. Google Translate,
4. Bing Translator,
5. MOSES statistical MT toolkit (Koehn et al., 2007).

The Google translate and the Bing translator are statistical MT systems. The main advantage of these two systems is that these are trained on huge corpora. Thus, the commonly used phrases will be translated in high quality, but in the case of unseen or rare segments or word forms, the quality will be low. In contrast, the MetaMorpho rule based MT system can handle numerous grammatical forms. Thus, it can gain high quality both in adequacy and fluency. The MOSES MT toolkit was trained on Hunglish corpus, which contains $\sim$1.1 million English-Hungarian sentence pairs, which is not big enough to perform high quality translation. There is a typical difference between statistical based MT systems and rule based MT systems for English-Hungarian. In the Table 1 we can see an example: *Smith turned the question over in his mind.* The main problem is that not Smith turned over, but the question turned over (by Smith).

[2] http://nlpg.itk.ppke.hu/node/65

MT system	Example	Adequacy			Fluency		
		D	L	Y	D	L	Y
Source	Smith turned the question over in his mind.						
MetaMorpho	Smith a kérdést forgatta a fejében.	2	5	5	4	5	5
Google	Smith megfordult a kérdés felett a fejében.	1	3	5	5	3	4
Bing	Smith megfordult a kérdés a fejében.	4	5	4	4	4	4
MOSES	Cyrus smith a kérdést.	1	1	1	1	1	4

Table 1: Example of translation difference

Adequacy	Fluency
1: none	1: incomprehensible
2: little meaning	2: disfluent Hungarian
3: much meaning	3: non-native Hungarian
4: most meaning	4: good Hungarian
5: all meaning	5: flawless Hungarian
0: I do not understand this English sentence	

Table 2: Adequacy and fluency scales for human evaluation

MetaMorpho using the grammatical analyzer could handle this problem correctly, but the statistical systems could not, because the probability of "Smith turning over" is higher than a "question turning over". This problem appears in the human evaluation as well. We can see in Table 1, in the case of Google translation, that the 3 annotators gave 3 different scores. One reason of the difference is that the 3 annotators had different kinds of point of view, another reason is the ambiguity. If we translate the Hungarian sentence back, it means: *Smith turned around in his mind, above the question*. In this case D gave 1 because this translation is totally different from the source sentence. But Y gave 5, because these phrases: "in mind" , "turn question" , together definitely have the main meaning that Smith analyzed the question, which has the same meaning as the source sentence. L agrees with both D and Y, he is halfway between D and Y.

For building the QE models, we used the arithmetic mean of the 3 annotators:

- AD: arithmetic mean of the adequacy scores,
- FL: arithmetic mean of the fluency scores,
- AF: arithmetic mean of the AD and FL scores.

We also created classification scores, because there are many cases, when we do not need 5 grades. For instance, the companies and translators need only 2 or 3 classes: need post-edit – do not need post edit; correct – need correction, etc. We created 3 classes from the AD, FL and AF scores:

- BAD: $1 \leq x \;\&\; x \leq 2$,
- MEDIUM: $2 < x \;\&\; x < 4$,
- GOOD: $4 \leq x \;\&\; x \leq 5$,

where: $x = AD, FL, AF$. The classification scores are:

- CLAD: classification scores from AD,
- CLFL: classification scores from FL,
- CLAF: classification scores from AF.

5. Methods, experiments and optimization

Using the HuQ corpus with AD, FL, AF, CLAD, CLFL and CLAF, we built the QE models. For building the QE model, features as quality indicators are needed to be extracted from the corpora. Then, with a machine learning method, human or automatic evaluation scores are used to build the QE model. To create the quality indicators from features, we used the QuEst framework. In this study, 103 features (103F) were extracted from the corpora. The set of 103 features contains 76 features implemented by Specia et al. and 27 additional features developed by us. In the 103F, there are adequacy features (e.g. ratio of percentage of nouns in the source and target, ratio of number of tokens in source and target, etc.), fluency features (e.g. perplexity of the target, percentage of verbs in the target, etc.) and complexity features (e.g. average source token length, source sentence log probability, etc.). The 27F contains 3 dictionary features and 24 features using WordNet and word embedding models.

The first task was doing evaluations with differently-sized portions of the HuQ corpus. Secondly, we evaluated the HuQ corpus with standard automatic metrics. Thereafter, we built different QE models for English-Hungarian. First, we tried the 17 baseline feature set (17F) (Specia et al., 2013) for Hungarian. The 17F is language and language tool independent. Then we performed experiments with the 103F (17F is subset of 103F). The problem was that the 103F contains features that use language dependent linguistic tools (e.g. Stanford parser (De Marneffe et al., 2006), Berkeley Parser (Petrov et al., 2006) etc.). The most commonly used linguistic tools could not be used for Hungarian. Thus, we integrated the available Hungarian linguistic tools into QuEst: For Part-of-Speech (POS) tagging and lemmatization, we used PurePos 2.0 (Orosz and Novák, 2013), which is an open source, HMM-based morphological disambiguation tool. Purepos2 has the state-of-the-art performance for Hungarian. It has the possibility to integrate a morphological analyzer. Thus, to get the best performance, we used Humor (Prószéky, 1994), a Hungarian morphological analyzer. For NP-chunking, we used HunTag (Recski and Varga, 2009) that was trained on the Szeged Treebank (Csendes et al., 2005). HunTag is a maximum entropy Markov-model based sequential tagger.

There are many language specific features that could not be extracted, because there are no Hungarian language tools for them.

For the machine learning task, we used the Weka system (Hall et al., 2009). We created 7 classifiers with 10 fold cross-validation: Gausian Processes with RBF kernel, Support Vector Machine for regression with NormalizedPolyKernel (SMOreg), Bagging (with M5P classifier), Linear regression, M5Rules, M5P Tree and for classification we used Support Vector Machine with NormalizedPolyKernel (SMO). Further on, we show only the results of the SMOreg and SMO, because these methods gained the best results. For evaluating the performance of our methods, we used the statistical correlation, the MAE (Mean absolute error), the RMSE (Root mean-squared error) and the Correctly Classified Instances (CCI) evaluation metrics. The correlation ranges from -1 to +1, and the closer the correlation to -1 or +1, the better it is. In the case of MAE and RMSE the closer the value to 0, the better.

We developed 27 new word-level semantic features. Our aim was to quantify the similarity and relatedness of the topic or meaning of the source and the target sentences. We collected only the *nouns*, *verbs*, *adjectives* and *adverbs* from the sentences. We created bag of words (BOW) from the source and the target segments. The bag of words contains the stem, the synonym and the semantic neighbors of the words.

There are 3 features extracted from an English-Hungarian dictionary used by MetaMorpho, which contains 365000 entries. We created noun, verb, adjective BOW from the source and the target sentences, then we counted the source-target word pairs from the BOW, which are contained by the dictionary. After all, we divided the matches by the length of the source sentence, the length of the target sentence and we counted the F1 score of them.

We developed 24 features using WordNet and word embedding models. We used the Princeton WordNet 3.0 (Fellbaum, 1998) and the Hungarian WordNet (Miháltz et al., 2008). We collected the synsets of the words in the source and the target segments. Then, we collected the hypernyms of the synsets up to two levels. Using the collected synsets and hypernym synsets we counted the weighted intersection of synsets of the source and the target words. Features are extracted from the result synsets.

However, if looking up words in WordNet did not provide any results, which is quite often the case because of the small coverage of the Hungarian WordNet, we used word embedding models to substitute synset results. Thus, first we trained a CBOW model with 300 dimensions on a 3-billion-word lemmatized Hungarian corpus. The reason for using the lemmatized version was to have semantic relations between words, rather than syntactic ones. Due to the agglutinating behaviour of Hungarian, building an embedding model from the raw text would have provided syntactically similar groups of words, and only a second key of similarity would have been their semantic relatedness. However, in the lemmatized model, this problem was eliminated. Thus, if there was no result for a word from WordNet, its top 10 nearest neighbours were retrieved from this embedding model and used the same way as WordNet synsets. However, as these lists do not necessarily correspond to synonyms of the original word, the weight of this feature was lower: weight = 0.1.

We carried out experiments for five different settings:

1. task (T1): we did statistical and inter-annotator agreement measurements on HuQ.
2. task (T2): we evaluated the quality of MT systems.
3. task (T3): HuQ corpus is evaluated using automatic evaluation methods: TER, BLEU and NIST (Lin and Och, 2004)
4. task (T4): using HuQ corpus and the 103F, we built QE models with different size of HuQ corpus trained on AF: 100, 500, 1000 and 1500 sentence pairs.
5. task (T5): using HuQ corpus, the 17F and the 103F, we built QE models trained on the automatic evaluation metrics, the AD, the FL and the AF scores.
6. task (T6): using HuQ corpus and the optimized feature sets, we built the QE models trained on the AD, the FL, the AF, the CLAD, the CLFL and the CLAF scores.

The experiment with human scores needed to be optimized for English-Hungarian. For optimizing, we used the forward selection method. First, we extracted and evaluated each feature separately. Then we chose the feature that produced the best result. Thereafter, we combined the chosen feature with each remaining feature, and we added the feature that produced the best combined result in each round. Then, we continued adding features until the combined result did not improve further.

6. Results and Evaluation

During T1, in Table 3 we can see the inter-rater agreement and in Figure 1 we can see the marginal distributions. Because of the ambiguities described in Section 4, the Fleiss Kappa values of inter-annotator agreement between the 3 annotators is moderate.

During T2 (see Table 4), as we expected, MOSES achieved the poorest result and MetaMorpho performed best.

The results of T3 describe the quality of the HuQ corpus. The system-level result of the T3 evaluation: TER: 0.6107; BLEU: 0.3038, NIST: 5.1359. According to the TER and the BLEU scores, ∼30% of the HuQ corpus are correct translations.

According to CLAF scores, we counted GOOD classes, there are 780 instances of GOOD, which means 52% of HuQ corpus are correct or close to correct translations. According to AF scores, we counted the 5 scores, there are 387 instances of 5 score, which means 25.8% of HuQ are correct translations.

During T4, as we can see in Table 5, increasing the size of HuQ, we got better results:

- the AF-500 could gain ∼24% higher correlation than the AF-100,
- the AF-1000 could gain ∼3% higher correlation than the AF-500,

	AD	FL	CLAD	CLFL	CLAF
Fleiss Kappa	0.357	0.463	0.44	0.521	0.493
Krippendorff Alpha	0.357	0.463	0.44	0.521	0.493
Average Pairwise Cohen's Kappa	0.360	0.464	0.444	0.522	0.494
Average pairwise percent	52.467%	61.222%	70.022%	74.444%	70.6%

Table 3: Evaluation of annotator-rater agreement

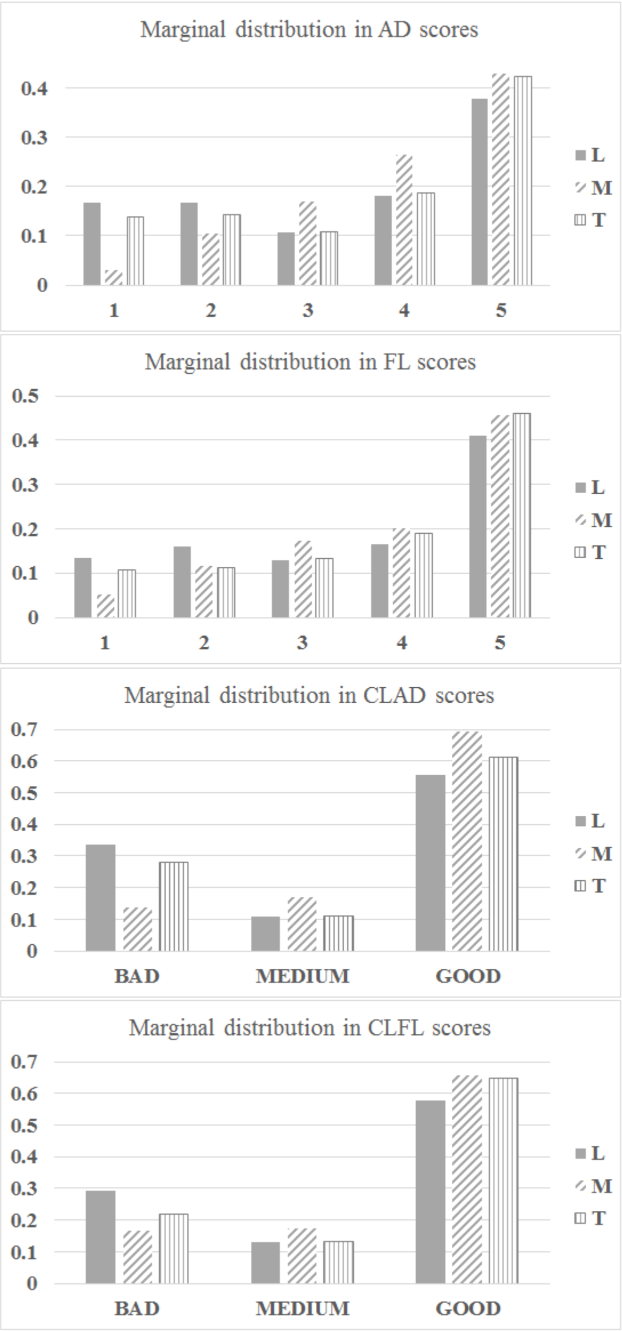

Figure 1: Marginal distributions

- the AF-1500 could gain ∼1.5% higher correlation than the AF-1000.

	AD mean	FL mean	AF mean
MetaMorpho	3.8707	3.8651	3.8679
MOSES	3.0175	3.1872	3.1024
Google	3.6395	3.5729	3.6062
Bing	3.2166	3.2256	3.2211

Table 4: Quality of MT systems

	Correlation	MAE	RMSE
AF-100	0.2700	0.8159	1.0613
AF-500	0.5155	0.8478	1.0603
AF-1000	0.5480	0.8147	1.0481
AF-1500	**0.5618**	**0.7962**	**1.0252**

Table 5: Evaluation of T4

During the results of T5 and T6 experiments, we built the QE models to predict the standard automatic evaluations and the human judgements. As we can see in Table 7, the AD-103F could gain ∼10% higher correlation than the 17F baseline set, the FL-103F could gain ∼6% higher correlation than the 17F baseline set, the AF-103F could gain ∼7% higher correlation than the 17F baseline set.

During T6, first, we used the 103F to build QE models trained on AD, FL and AF human scores. Then, we optimized the models to Hungarian. After optimizing, as we can see the results in Table 7, the optimized AD set containing 29 features could gain ∼4% higher correlation than the 103F and ∼14% higher correlation than the 17F baseline set, the optimized FL set containing 32 features could gain ∼4% higher correlation than the 103F and ∼10% higher correlation than the 17F baseline set, the optimized AF set containing 26 features could gain ∼5% higher correlation than the 103F and ∼12% higher correlation than the 17F baseline set.

Then, we did experiment, evaluation and optimization with the classification scores. As we can see in Table 8, the optimized CLAD set containing 21 features could gain ∼3% higher correlation than the 103F and ∼6% higher correlation than the 17F baseline set, the optimized CLFL set containing 10 features could gain ∼1.5% higher correlation than the 103F and ∼5% higher correlation than the 17F baseline set and the optimized CLAF set containing 12 fea-

	Correlation	MAE	RMSE
TER	0.3550	0.3275	0.4357
BLEU	0.4404	0.2201	0.3474
NIST	0.3669	2.6695	3.4777

Table 6: Evaluation of QE using the standard metrics

	Correlation	MAE	RMSE
AD-17F	0.3832	0.9429	1.1990
FL-17F	0.5400	0.8229	0.8345
AF-17F	0.4931	0.8345	1.0848
AD-103F	0.4847	0.8805	1.1199
FL-103F	0.6070	0.7723	1.0297
AF-103F	0.5618	0.7962	1.0252
Optimized AD	**0.5245**	**0.8397**	**1.0869**
Optimized FL	**0.6413**	**0.7440**	**0.9878**
Optimized AF	**0.6100**	**0.7459**	**0.9775**

Table 7: Evaluation QE using the human judgements

tures could gain $\sim$1.5% higher correlation than the 103F and $\sim$4% higher correlation than the 17F baseline set.

	CCI	MAE	RMSE
CLAD-17F	54.9333%	0.3590	0.4591
CLFL-17F	58.8667%	0.3434	0.4419
CLAF-17F	57.8000%	0.3433	0.4417
CLAD-103F	57.6667%	0.3492	0.4483
CLFL-103F	62.4667%	0.3310	0.4275
CLAF-103F	60.3333%	0.3347	0.4318
Optimized CLAD	**60.9333%**	**0.3370**	**0.4346**
Optimized CLFL	**64.0667%**	**0.3299**	**0.4262**
Optimized CLAF	**61.8000%**	**0.3299**	**0.4263**

Table 8: Evaluation of QE using the classification metrics

7. Conclusion

We created the HuQ corpus for quality estimation of English-Hungarian machine translation. The corpus contains 1500 quality scores of translations, which are given by human annotators. Then using the HuQ corpus, we built different QE models for English-Hungarian translations. In our experiments, we used automatic metrics and human judgements as well. In the experiments we tried 103 features including 27 newly developed semantic features using WordNet and word embedding models. Then, we optimized the quality models to English-Hungarian. In the optimization task, we used forward selection to find the best features. We could produce optimized sorted feature sets, which produced more than 10% better correlation than the baseline set. In our experiments, our HuQ corpus and QE models can be used for predicting the quality of machine translation outputs for English-Hungarian.

In the future, we would like to enlarge the corpus. We also would like to examine the effect of utilizing crowd sourcing to increase the size of HuQ. Last, but not least, We will do experiments and evaluations in ranking task.

Acknowledgements

The authors would like to thank Andrea Dömötör and László Laki for their generous help in the annotation work.

References

Beck, D., Shah, K., Cohn, T., and Specia, L. (2013). Sheflite: When less is more for translation quality estimation. In *Proceedings of the Workshop on Machine Translation (WMT)*, August.

Biçici, E. (2013). Feature decay algorithms for fast deployment of accurate statistical machine translation systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Camargo de Souza, J. G., Buck, C., Turchi, M., and Negri, M. (2013). FBK-UEdin participation to the WMT13 quality estimation shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 352–358, Sofia, Bulgaria, August. Association for Computational Linguistics.

Csendes, D., Csirik, J., Gyimóthy, T., and Kocsor, A. (2005). The Szeged Treebank. In *Lecture Notes in Computer Science: Text, Speech and Dialogue*, pages 123–131. Springer.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Halácsy, P., Kornai, A., Németh, L., Sas, B., Varga, D., Váradi, T., and Vonyó, A. (2005). A Hunglish korpusz és szótár. In *III. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Egyetem.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

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., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., and Váradi, T. (2008). Methods and results of the hungarian wordnet project. In *Proceedings of the Fourth Global WordNet Conference GWC 2008*, pages 310–320.

Novák, A., Tihanyi, L., and Prószéky, G. (2008). The metamorpho translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 111–114, Stroudsburg, PA, USA. Association for Computational Linguistics.

Orosz, G. and Novák, A. (2013). Purepos 2.0: a hybrid tool for morphological disambiguation. In *RANLP'13*, pages 539–545.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.

Prószéky, G. (1994). Industrial applications of unification morphology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 213–214, Stuttgart, Germany, October. Association for Computational Linguistics.

Recski, G. and Varga, D. (2009). A Hungarian NP Chunker. *The Odd Yearbook. ELTE SEAS Undergraduate Papers in Linguistics*, pages 87–93.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Specia, L., Shah, K., de Souza, J. G., and Cohn, T. (2013). Quest - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria, August. Association for Computational Linguistics.