Quality Estimation for English-Hungarian Machine Translation

Zijian Győző Yang[*], **László János Laki**[†]

[*]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}@itk.ppke.hu

Abstract

Quality estimation for machine translation systems has become an important task. There are automatic evaluation methods for machine translation that use reference translations created by human translators. The creation of these reference translations is very expensive and time-consuming. Furthermore, these automatic evaluation methods are not real-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. The other kind of evaluation approach is quality estimation. These methods address the task by estimating the quality of translations as a prediction task for which features are extracted from only the source and translated sentences. This approach has not been used for evaluating English to Hungarian translations before. In this study, a corpus is created, which contains human judgements. Using these human evaluated scores, different quality estimation models are described. These models are evaluated and optimized based on human scores for English-Hungarian translations. We developed 27 new features and created a feature set, which produced better results than the baseline feature set for Hungarian.

1. Introduction

As machine translation (MT) has become popular among people and companies, the measurement of the translation output has become necessary. A quality indicator for MT could save a lot of time and money for users. Knowing the quality scores of machine translated segments can help human annotators in their post-edit tasks, or help MT systems to find and combine the best translations. Last but not least, quality indicators can filter out and inform about unreliable translations.

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. These methods are automatic evaluation approaches such as BLEU, and other methods based on BLEU, TER, HTER etc. The problem is that automatic evaluation methods cannot perform well enough in this task, because these need reference translations. It means that after the 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 very expensive and slow. Thus, 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 does not use reference translations. This supervised approach is called Quality Estimation (QE) of MT. 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 human annotators, researchers and companies.

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.

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. Thus, 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 quality estimation approach. Then, we will present our experiments in the task of quality estimation and the optimization of these models are described.

2. Related Work

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 indicators. Then the QE model is used to predict the quality of unseen translations. The aim is that the QE model correlates with human judgments, thus the QE model is trained on human judgments. In the last couple of years there have been many WMT workshops with quality estimation shared tasks,[1] which provided datasets for QE researches. The datasets are evaluated with HTER, METEOR, ranking or post-editing effort scores. But unfortunately there is no dataset for Hungarian. In this research we created a dataset for Hungarian and for human judgement we used a general scoring scale.

Recently, 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

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

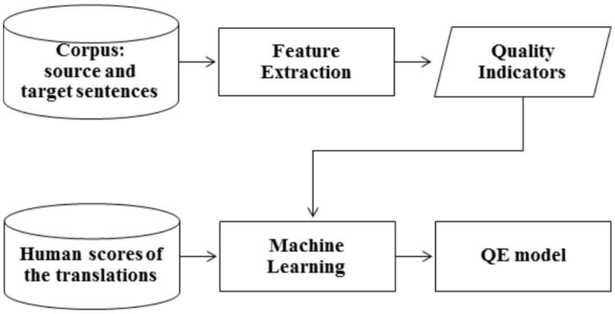
Corpus: source and target sentences
Feature Extraction
Quality Indicators
Human scores of the translations
Machine Learning
QE model

Figure 1: QE algorithm

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). There is a language independent baseline set which contains 17 features. In our research we did experiments for Hungarian in both fields.

3. Quality Estimation

In the quality estimation task (see Fig. 1), using various features, we extract different kinds of quality indicators from the source and translated sentences without using reference translations. From the source sentences, complexity features can be extracted (e.g. number of tokens in the source segment). From the translated sentences, QuEst extracts 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 features from the decoder of the MT system. These are the confidence features (e.g. features and global score of the SMT system). We can divide the features into two more 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 the feature extraction, using these quality indicators, we can build QE model with machine learning methods. The aim is that the predictions of the QE model are highly correlated with human evaluations. Thus, the extracted quality indicators need to be trained on human judgments.

4. Datasets

In our experiments, we used two corpora. The first corpus (C1) contains 1950 English sentences of mixed topics from the Hunglish corpus (Halácsy et al., 2005). Beside the human-translated sentence pairs in this corpus, each segment was also translated by Google translate, Bing translate, the MetaMorpho (Novák et al., 2008) rule based MT system and the MOSES statistical MT toolkit (Koehn et al., 2007). The reference sentences are human translated sentences from the Hunglish corpus. In order to get the best performance of the QE model prediction, this training corpus was created by selecting sentences from the Hunglish corpus, which had perfect human translations. The second corpus (C2) is a subset of the first one, which contains 550 segments with human evaluated scores.

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

Table 1: Adequacy and fluency scales for human evaluation

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 a translated Hungarian sentence. People can give quality scores from 1 to 5, from two points of view (Koehn, 2010): adequacy and fluency (see Table 1). We added a 0 score (*I do not understand the English sentence*) to filter out wrong evaluations. 550 sentences were evaluated by at least 3 human translators. The translators were Hungarian people who have minimum B2 level English language skill. For building the QE model, we used the mean of the Adequacy and the Fluency scores.

5. Methods and Experiments

For building the QE model, features as quality indicators are needed, which are extracted from the corpora. Then, with a machine learning method, human or automatic evaluation scores are used to build the QE model (see Figure 1). 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 (76F) implemented by Specia et al. and 27 additional features developed by us (27F). In the 76F, 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 WordNet features.

The first task was trying features developed by Specia et al (Specia et al., 2013). 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 76F (17F is subset of 76F). The problem was that the 76F 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

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

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.

5.1. Dictionary and WordNet

We used 3 features extracted from an English-Hungarian dictionary (Novák et al., 2008):

$$\frac{number\ of\ matches}{length\ of\ source\ sentence} \tag{1}$$

$$\frac{number\ of\ matches}{length\ of\ target\ sentence} \tag{2}$$

$$harmonic\ mean\ of\ (1)\ and\ (2) \tag{3}$$

We developed 24 features extracted from WordNet. 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 sentence. Then, we collected the hypernyms of the synsets in two levels. Using the collected synsets and hypernym synsets we counted the intersection of synsets of the source and the target words. Features are extracted from the result synsets:

$$\frac{number\ of\ weighted\ (x\ matches)}{length\ of\ source\ sentence} \tag{4}$$

$$\frac{number\ of\ weighted\ (x\ matches)}{number\ of\ x\ in\ source\ sentence} \tag{5}$$

$$\frac{number\ of\ weighted\ (x\ matches)}{length\ of\ target\ sentence} \tag{6}$$

$$\frac{number\ of\ weighted\ (x\ matches)}{number\ of\ x\ in\ target\ sentence} \tag{7}$$

$$harmonic\ mean\ of\ (4)\ and\ (6) \tag{8}$$

$$harmonic\ mean\ of\ (5)\ and\ (7) \tag{9}$$

where:

$$x = nouns,\ verbs,\ adjectives,\ adverbs$$

$$weighted\ (x\ matches) = \sum \frac{x\ match}{level}$$

TER	0.6107
BLEU	0.3038
NIST	5.1359

Table 2: Evaluation of T1

5.2. Machine learning

For the machine learning task, we used the Weka system (Hall et al., 2009) to create 6 classifiers with 10 fold cross-validation: Gausian Processes with RBF kernel (GP), Support vector machine for regression with NormalizedPolyKernel (SMOreg), Bagging (with M5P classifier), Linear regression, M5Rules and M5P Tree. Further on, we show only the results of the first two classifiers, because these methods gain the best scores. For evaluating the performance of our methods, we used the statistical correlation, the MAE (Mean absolute error) and the RMSE (Root mean-squared error) evaluation metrics. The correlation ranges from -1 to +1, and the closer the correlation to -1 or +1 is, the better it is. In the case of MAE and RMSE the closer the value to 0, the better.

5.3. Experiments and optimization

We carried out experiments for four different settings:

- First task (T1): C1 is evaluated using automatic evaluation methods: TER, BLEU and NIST(Lin and Och, 2004).
- Second task (T2): using the automatic evaluation metrics (segment-level TER, BLEU and NIST), the 103F and C1, QE models were built and evaluated.
- Third task (T3): using C2 and the 103F, we built and evaluated QE models trained on adequacy scores, fluency scores and mean score of adequacy and fluency scores (A+F).
- Fourth task (T4): using C2, 17F, 76F and 103F, we built the QE model, then optimized the 103F for Hungarian.

The experiment with human scores needed to be optimized for English-Hungarian. For optimizing, we used the forward selection method. First, we separately extracted and evaluated each feature. 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. With this attribute selection algorithm, we could create a feature set that contains 23 features (23F).

6. Results and Evaluation

The results of T1 describe the quality of the C1 corpus. According to the TER and the BLEU scores, 30% of the C1 corpus are correct translations. The system-level result of the T1 evaluation can be found in Table 2.

To predict the automatic evaluation scores with QuEst, as we can see in Table 3, the GP method achieved the best results in TER evaluation, in all cases. In BLEU and NIST evaluations, the SMOreg method won in correlation and MAE scores, but in the RMSE, the GP method was the best.

		TER	BLEU	NIST
GP	Corr	0.3672	0.4028	0.3254
	MAE	0.3202	0.2598	2.7680
	RMSE	0.4277	0.3335	3.4438
SMOreg	Corr	0.3550	0.4404	0.3669
	MAE	0.3275	0.2201	2.6695
	RMSE	0.4357	0.3474	3.4777

Table 3: Evaluation of T2

In the T3 experiment, the SMOreg methods gained the best results. As we can see in Table 4, the models trained on Fluency scores achieved the best results. It is interesting to note that during the SMOreg experiment, the features having the highest weights, are mostly complexity features.

		Adequacy	Fluency	A+F
GP	Corr	0.4934	0.5705	0.5536
	MAE	1.0347	0.9407	0.9279
	RMSE	1.1975	1.1208	1.0952
SMOreg	Corr	0.5058	0.6147	0.5851
	MAE	0.9642	0.8514	0.8621
	RMSE	1.2064	1.0827	1.0739

Table 4: Evaluation of T3

During T4, first, we used the 103F and A+F human scores to build and evaluate the QE models. Then, using the forward selection method, 23 features were selected, and using these 23F, we built the QE models and evaluated them. We also evaluated the C2 with the 17F and 76F. As we can see in Table 5, using SMOreg, with the 103F we could achieve ˜1% better correlation than with the 76F and ˜7% better correlation than with the 17F. Using our forward selection method, we could gain ˜11% better correlation than with the 17 baseline features set. Even more the 23F could produce ˜4% higher correlation than the 103F.

		Corr	MAE	RMSE
17F	GP	0.5101	0.9333	1.1217
	SMOreg	0.5112	0.912	1.1353
76F	GP	0.5763	0.9076	1.0925
	SMOreg	0.5784	0.9036	1.1214
103F	GP	0.5536	0.9279	1.0952
	SMOreg	0.5851	0.8621	1.0739
23F	GP	0.5859	0.8704	1.0578
	SMOreg	**0.6275**	**0.795**	**1.0292**

Table 5: QE model optimization for Hungarian (T4)

In Table 6, we can see the 23 features for Hungarian. The 3 features in bold has just been developed in this research.

7. Conclusion

First, we created a training corpus for English-Hungarian translation quality estimation. The corpus contains quality scores of translations, which are evaluated by human translators. Then using the human judgements, we built different QE models for English-Hungarian translations. In our experiments, we used automatic metrics and human judgements as well. In the experiment we tried 103 features which contains 27 newly developed features. Then, we optimized the quality model to English-Hungarian. In the optimization task, we used the forward selection to find the best features. We could produce a sorted feature set which contains 23 features. These 23 features produced 11% higher correlation than the 17F baseline feature set, furthermore this 23F produced better results than the 103F. We propose that these 23 features can be the baseline set for English-Hungarian translations. In our experiments, we built different QE models, which can be used for predicting the quality of machine translation outputs for English-Hungarian.

In the future, we plan to try more features implemented by Specia at al. and also further investigate the addition of our own features. We also plan to increase the size of the C2 corpus. Our goal is to build a stable and reliable QE model for English-Hungarian.

8. References

Beck, Daniel, Kashif Shah, Trevor Cohn, and Lucia Specia, 2013. Shef-lite: When less is more for translation quality estimation. In *Proceedings of the Workshop on Machine Translation (WMT)*.

Biçici, Ergun, 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: Association for Computational Linguistics.

Camargo de Souza, José Guilherme, Christian Buck, Marco Turchi, and Matteo Negri, 2013. FBK-UEdin participation to the WMT13 quality estimation shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics.

Csendes, Dóra, János Csirik, Tibor Gyimóthy, and András Kocsor, 2005. The Szeged Treebank. In *Lecture Notes in Computer Science: Text, Speech and Dialogue*. Springer.

De Marneffe, Marie-Catherine, Bill MacCartney, Christopher D Manning, et al., 2006. Generating typed depen-

absolute difference between number of commas in source and target
percentage of tokens in the target which do not contain only a-z
percentage of verbs in the target
dictionary lookup f-score
percentage of verbs in the source
perplexity of the target
number of tokens in target
average bigram frequency in quartile 2 of frequency in the corpus of the source sentence
dictionary lookup precision
source sentence perplexity
percentage of punctuation marks in target
average unigram frequency in quartile 1 of frequency in the corpus of the source sentence
absolute difference between number of : in source and target
WordNet count in source: nouns / number of nouns
average unigram frequency in quartile 2 of frequency in the corpus of the source sentence
ratio of percentage of tokens a-z in the source and tokens a-z in the target
average trigram frequency in quartile 1 of frequency in the corpus of the source sentence
source sentence perplexity without end of sentence marker
absolute difference between number of ! in source and target normalized by target length
absolute difference between number of ! in source and target
average bigram frequency in quartile 3 of frequency in the corpus of the source sentence
absolute difference between number of : in source and target normalised by target length
number source tokens that do not contain only a-z

Table 6: 23 features for Hungarian

dency parses from phrase structure parses. In *Proceedings of LREC*, volume 6.

Fellbaum, Christiane, 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

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

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Koehn, Philipp, 2010. *Statistical Machine Translation*. New York, NY, USA: Cambridge University Press, 1st edition.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst, 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. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, Chin-Yew and Franz Josef Och, 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, Márton, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi, 2008. Methods and results of the hungarian wordnet project. In *Proceedings of the Fourth Global WordNet Conference GWC 2008*.

Novák, Attila, László Tihanyi, and Gábor Prószéky, 2008. The metamorpho translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08. Stroudsburg, PA, USA: Association for Computational Linguistics.

Orosz, György and Attila Novák, 2013. Purepos 2.0: a hybrid tool for morphological disambiguation. In *RANLP'13*.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein, 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*. Association for Computational Linguistics.

Prószéky, Gábor, 1994. Industrial applications of unification morphology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*. Stuttgart, Germany: Association for Computational Linguistics.

Recski, Gábor and Dániel Varga, 2009. A Hungarian NP Chunker. *The Odd Yearbook. ELTE SEAS Undergraduate Papers in Linguistics*:87–93.

Specia, Lucia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn, 2013. Quest - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria: Association for Computational Linguistics.