eπQue: The Machine Translation Quality Estimation Software Package

Doctor of Philosophy Dissertation

Zijian Győző Yang

Roska Tamás Doctoral School of Sciences and Technology
Pázmány Péter Catholic University
Faculty of Information Technology and Bionics

Academic Supervisor
Gábor Prószéky, DSc

Budapest, 2019

Chapter 1

Introduction

Machine translation has become a daily used tool among people and companies. The measurement of the quality of translation output has become necessary. A quality score for machine translation could save a lot of time and money for users, companies and researchers. Knowing the quality of machine translated segments can help human annotators in their post-edit tasks, or using the quality we can filter out and inform users about unreliable translations. Last but not least, quality indicators can help machine translation systems to combine the translations to produce better output.

There are two kinds of evaluation methods for machine translation. The first type is the manual human evaluation, which is the most expensive, slowest and subjective method, but is still the most accurate. The other type is the automatic machine evaluation. This kind of evaluation is less accurate, but faster and cheaper than the human evaluation. The machine evaluation is always based on human evaluation. We can separate the machine evaluation into two categories. The first category uses reference translations, i.e. it compares machine translated sentences to human translated reference sentences, then it measures the similarities or differences between them. The problem is that automatic evaluation methods 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 expensive and time-consuming. We can not use these methods in run-time. The other type of category is a prediction method, which is called quality estimation. The quality estimation (QE) uses the source and the MT translated seg-

ments to extract different kinds of quality indicators, then using these quality indicators, a machine learning algorithm is trained. The QE can evaluate segments in real-time and does not need reference translations.

In my Thesis, I used the quality estimation method in three different tasks. First, I implemented the quality estimation method in English-Hungarian. In this task, I created a human annotated corpus to train the English-Hungarian quality estimation model. In order to build a model, features are required. I implemented quality estimation features, that optimized for other languages, then I created new semantic features for English-Hungarian. Further optimizations were performed as well.

Secondly, using my English-Hungarian quality estimation system I combined different machine translation outputs to achieve higher translation quality at the system level. I created a composite machine translation system that could gain better quality than the systems are used by the composite system. I tested my method in more different language pair and I could produced better result in all cases.

Finally, I used the quality estimation method to predict the quality of monolingual texts. Using the quality estimation algorithm, I built a task-oriented quality estimation module to detect the quality and error types of a monolingual text.

The eπQue quality estimation software package consists of these three applications described above.

Chapter 2

New scientific results

In my research, I used the quality estimation method in three research fields. First, I created an English-Hungarian quality estimation system (Hun-QuEst system). To train the system, I built a human manual evaluated corpus (HuQ corpus). Furthermore, I created 27 new semantic features, which produced higher results than the baseline feature set. In my second task, I used the quality estimation technique to combine different kinds of machine translation system. I built a composite system (MaTros system), that combine machine translation outputs to achieve higher translation quality at the system level. Finally, I used the quality estimation method to predict quality and detect errors in monolingual text (πRate system).

The eπQue software package contains the three system, that I described before (Hun-QuEst system; MaTros system; πRate system).

2.1 English-Hungarian quality estimation

The quality estimation method is based on machine learning. The model (See Figure 2.1.) extract features as quality indicators from source and machine translated segments. Then, using a machine learning algorithm and the quality indicators, the quality estimation model is trained on human evaluations.

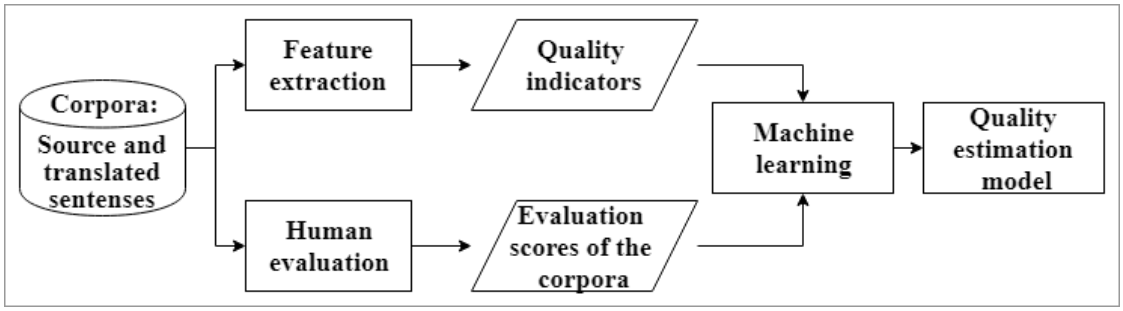

Figure 2.1 Architecture of the quality estimation model

To train the quality estimation model, training corpora are needed. But unfortunately there are no human evaluated parallel corpora for English-Hungarian.

Thesis 1: **I created a human evaluated corpus for English-Hungarian quality estimation system.**

Related publications: [6] [8].

The HuQ corpus contains 1500 English-Hungarian sentence pairs. To build the HuQ corpus, I used 300 English sentences of mixed topics from the Hunglish corpus. I translated these 300 sentences into Hungarian with different machine translation systems: MetaMorpho rule based machine translation system, Google Translate, Bing Translator and MOSES statistical machine translation toolkit. After the translation, to create human judgements, I evaluated these translated segments with human annotators. All the 1500 sentences were evaluated by 3 human annotators: a linguist, a machine translation specialist and a language technology expert. The annotators could give quality scores from 1 to 5, based on 2 evaluation criteria: adequacy and fluency. For a binary classification task, I created 2 class labels from the human judgement scores: "ER" (erroneous translation): $x \leq 4$ and "OK" (correct translation): $x > 4$. For another classification task, I created 3 class labels from the human judgement scores: "BAD" (unusable translation): $1 \leq x \leq 2$, "MEDIUM" (need correction): $2 < x < 4$ és "GOOD" (usable translation): $4 \leq x \leq 5$.

Using the HuQ corpus, I built an English-Hungarian quality estimation system.

I did different experiments with the English-Hungarian quality estimation system. First, I tried the baseline features, that were optimized for English-Spanish. Then, I tested 76 features, which are implemented by Specia at al. Thereafter, using an English-

Hungarian dictionary, WordNet, word embedding and latent semantic analysis methods, I created new semantic features. Last, I did feature selection task, which means, I could gain higher results with less features.

	Correlation ↑	MAE ↓	RMSE ↓
TG-17F (baseline)	0.4931	0.8345	1.0848
TG-103F	0.5618	0.7962	1.0252
OptTG (26 features)	**0.6100**	**0.7459**	**0.9775**

Table 2.1 Evaluation of the Hun-QuEst regression models

	CCI ↑	MAE ↓	RMSE ↓
CLTG-17F (baseline)	57.8000%	0.3433	0.4417
CLTG-103F	60.3333%	0.3347	0.5495
OptCLTG (12 features)	**61.8000%**	**0.3299**	**0.4263**

Table 2.2 Evaluation of the Hun-QuEst classification models (3 class label)

	CCI ↑	MAE ↓	RMSE ↓
CLBITG-17F (baseline)	65.7333%	0.3427	0.5854
CLBITG-103F	69.7333%	0.3027	0.5502
OptCLBITG (16 features)	**70.1333%**	**0.2987**	**0.5465**

Table 2.3 Evaluation of the Hun-QuEst binary classification models

In Table 2.1., Table 2.2. and Table 2.3., we can see the evaluation of the Hun-Quest models.

I also tried the WordNet features for English-Spanish and English-German language pairs. I could gain better results than the baseline feature set in both cases.

Thesis 2: **Using a bilingual dictionary, the WordNet and the word embedding method, I created 27 new semantic features, that gained higher results than the baseline feature set.**

Thesis 3: **Using the English-Hungarian training corpus and the 27 semantic features, that I created, I built an English-Hungarian quality estimation system based on QuEst framework with integration of Hungarian linguistic tools.**

Related publications: [4] [6] [7] [9] [10].

2.2 Combining machine translation systems with quality estimation

In this research, I combined outputs of different machine translation systems.

Thesis 4: **Using the quality estimation method, I built a composite machine translation system, that combines outputs of different machine translation systems. The quality of the composite system is higher than the combined machine translation systems alone.**

Related publications: [5] [6] [12].

In this research I did the experiments in a business environment. The composite system (See Figure 2.2) combines outputs of a phrase-based statistical, a hierarchical-based statistical and a neural machine translation. Using the quality estimation method, the system chooses the translation that has the highest quality. The chosen translation will be the output of the composite system.

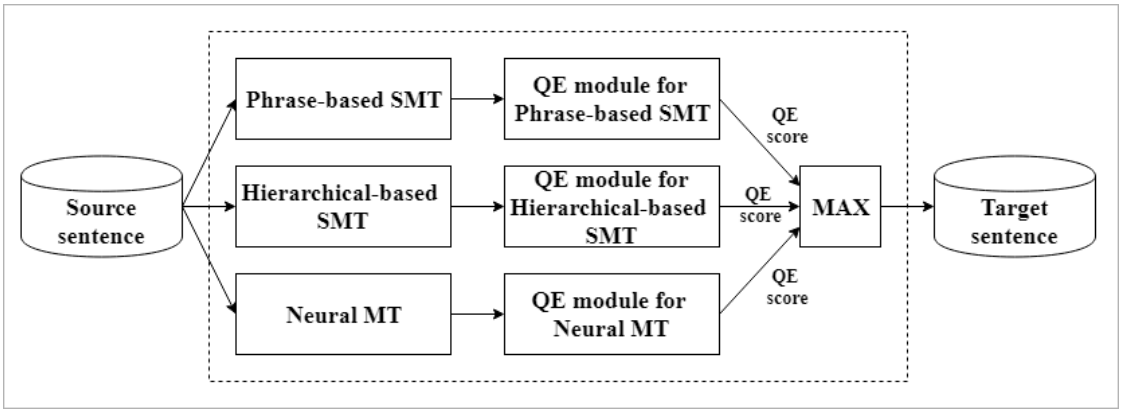

Figure 2.2 Architecture of the composite system

		en-hu	en-hu+	en-de	en-it	en-ja
BLEU mean ↑	PBSMT	0.5156		0.6288	0.7513	0.5945
	HBSMT	0.6157		0.4808	0.6998	0.6044
	NMT	0.6281		0.4364	-	-
	CoMT	**0.6926**	**0.6978**	**0.6662**	**0.7525**	**0.6057**
oBLEU mean ↑	PBSMT	0.7381		0.6757	0.8202	0.5361
	HBSMT	0.7679		0.6221	0.7993	0.5536
	NMT	0.7252		0.6751	-	-
	CoMT	**0.7729**	**0.7734**	**0.6855**	**0.8246**	**0.5553**
oTER mean ↓	PBSMT	0.2903		0.3574	0.1669	0.4281
	HBSMT	0.2193		0.4170	0.1995	0.4075
	NMT	0.2101		0.2653	-	-
	CoMT	**0.1892**	**0.1871**	**0.2649**	**0.1662**	**0.4055**

Table 2.4 Performance of the composite system

I tested my composite method in four different language pairs: English-Hungarian, English-German, English-Italian and English-Japanese. At the system level, my composite system achieved the highest result in all cases. For English-Hungarian I also did optimization with language dependent features, which produced further improvement in results.

In Table 2.4., we can see the performance of the composite system.

2.3 Monolingual quality estimation system for error detection

In this research, I used the quality estimation method for a monolingual task. The aims of this research are to analyse the human produced errors in monolingual texts available online and using quality estimation method, to build a quality prediction and error detection software.

Thesis 5: **Using quality estimation method, I created a monolingual error detection system, that can predict quality and detect human produced errors in monolingual text.**

Related publications: [1] [6] [11] [13].

The human produced errors are different from errors made by machine translation systems. Thus, different training corpora and features are needed. Andrea Dömötör built the corpus that I used to train my quality estimation model.

	Correlation ↑	MAE ↓	RMSE ↓
LS model - 36 features	0.7712	0.7121	1.0047
OptLS model - 15 features	0.7777	0.7226	0.9625

Table 2.5 Evaluation of the regression models

	CCI ↑	MAE ↓	RMSE ↓
CS model - 36 features	64.48%	0.214	0.3171
OptCS model - 28 features	65.17%	0.2137	0.3167

Table 2.6 Evaluation of the classification models

In Table 2.5. and Table 2.6, we can see the evaluation of the models.

This research shows that unlike machine translations, human errors are mostly not grammatical errors. These errors rather caused by the writing habits of the Internet users, for instance missing accents or punctuation marks. My monolingual quality estimation system well suited for corpus linguistic tasks or as a module can provide help in the preprocessing task of a natural language parser.

Chapter 3

The author's publications

Journal publications

[1] **Zijian Győző Yang** and L. J. Laki, "πRate: A Task-oriented Monolingual Quality Estimation System", *International Journal of Computational Linguistics and Applications*, 2017, [Accepted paper].

[2] A. Dömötör and **Zijian Győző Yang**, "What's your style? Automatic genre identification with neural network", *International Journal of Computational Linguistics and Applications*, 2018, [Accepted paper].

[3] **Zijian Győző Yang**, "A Gépi fordítás és a neurális gépi fordítás", *Modern Nyelvoktatás*, vol. 24, no. 2–3, pp. 129–139, 2018.

Book chapters

[4] **Zijian Győző Yang**, L. J. Laki, and B. Siklósi, "Quality Estimation for English-Hungarian with Optimized Semantic Features", in *Computational Linguistics and Intelligent Text Processing*, Konya, Turkey, 2016.

[5] L. J. Laki and **Zijian Győző Yang**, "Combining Machine Translation Systems with Quality Estimation", in *Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary: Springer International Publishing, 2017, pp. 435–444, ISBN: 978-3-319-77116-8.

[6] **Zijian Győző Yang**, A. Dömötör, and L. J. Laki, "A Quality Estimation System for Hungarian", in *Human Language Technology. Challenges for Computer Science and Linguistic*, Poznań, Poland: Springer International Publishing, 2018, pp. 201–213, ISBN: 978-3-319-93782-3.

International conference proceedings

[7] **Zijian Győző Yang** and L. J. Laki, "Quality Estimation for English-Hungarian Machine Translation", in *7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland: Uniwersytet im. Adama Mickiewicza w Poznaniu, 2015, pp. 170–174.

[8] **Zijian Győző Yang**, L. J. Laki, and B. Siklósi, "HuQ: An English-Hungarian Corpus for Quality Estimation", in *Proceedings of the LREC 2016 Workshop - Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, (May 24, 2016), Portorož, Slovenia, 2016.

Hungarian conference proceedings

[9] **Zijian Győző Yang** and L. J. Laki, "Gépi fordítás minőségének becslése referencia nélküli módszerrel", in *XI. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 2015, pp. 3–13.

[10] **Zijian Győző Yang** and L. J. Laki, "Gépi fordítás minőségbecslésének optimalizálása kétnyelvű szótár és WordNet segítségével", in *XII. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 2016, pp. 37–46.

[11] **Zijian Győző Yang** and L. J. Laki, "Minőségbecslő rendszer egynyelvű természetes nyelvi elemzőhöz", in *XIII. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 2017, pp. 37–49.

[12] L. J. Laki and **Zijian Győző Yang**, “Gépi fordító rendszerek kombinálása minőségbecslés segítségével”, in *XIV. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 2018, pp. 281–291.

[13] A. Dömötör and **Zijian Győző Yang**, “Így írtok ti - Nem sztenderd szövegek hibatípusainak detektálása gépi tanulásos módszerrel”, in *XIV. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 2018, pp. 305–316.

[14] **Zijian Győző Yang**, “A gépi fordítás kiértékelése”, in *Fókuszban a fordítás értékelése*, Budapest, Hungary: Budapesti Műszaki és Gazdaságtudományi Egyetem Gazdaság - és Társadalom tudományi Kar Idegen Nyelvi Központ, 2018, pp. 147–162.