FBK HLT-MT Participation in the 1^{st} Translation Memory Cleaning Shared Task

Duygu Ataman[1,2], Masoud Jalili Sabet[3], Marco Turchi[2], Matteo Negri[2]
[1] University of Trento, [2] Fondazione Bruno Kessler, [3] University of Tehran
{ataman,turchi,negri}@fbk.eu, jalili.masoud@ut.ac.ir

Abstract

We present the translation memory cleaning system designed by FBK HLT-MT to participate in the shared task of the NLP4TM 2016 workshop. Our system integrates different feature extraction approaches based on multilingual distributional semantics, machine translation quality estimation methods and word alignments in a supervised machine learning scheme. In order to assign a class to an input sentence pair, the system combines these features and feeds to an ensemble decision tree classifier. We compare our systems in terms of the effects of each set of features, different classifier settings and discuss the improvement on performance due to feature selection and sample weighting. Our system achieves the 1^{st} rank in the English-Spanish Binary-I classification task.

Keywords: translation memory, natural language processing, distributional semantics, word alignment, machine learning

1. Introduction

The 1st Translation Memory Cleaning Shared Task of NLP4TM 2016 is a novel workshop which aims to provide computational means to improve translation memories. Being one of the most important components of computer-assisted translation (CAT) tools by translators, translation memories lack advanced measures for maintenance and specialization of the translation units with respect to increased usage and time. This shared task thus addresses an extreme need of the translation community by providing a new framework in development and comparison of natural language processing methods, evaluation of machine translation output and integration of these measures into the translation memories.

The task is defined as a classification problem where a set of translation memory units (TU), each consisting of a reference sentence and its machine-translated version, are labelled according to the post-editing effort required for each sentence. The aim of the shared task is to develop an autonomous system which can learn to annotate new TUs according to the provided example set. Each TU may have one of the three labels corresponding to (1) correct translation, (2) almost correct translation with few syntactic or semantic flaws, (3) incorrect translation. The shared task has been organized into three sub-tasks according to the different perspectives adopted while defining the usefulness of a TU. The 1st and 2nd sub-tasks are binary classification tasks where the *True* label may be used only for correct translations (i.e. category '1') or may represent any translation output that is useful for the translator whether with few or none post-editing effort (i.e. categories '1' and '2').

In this paper, we present the system developed by FBK HLT-MT to participate in the NLP4TM 2016 shared task. We take advantage of this shared task to evaluate the performance of the current version of the Translation Memory Open-source Purifier (TMop) developed at FBK (Jalili et al., 2016) and improve the tool in terms of adequacy features to develop a new translation memory cleaning system. Our system is based on combining different approaches based on language-specific and syntactic information and semantic cues obtained by bilingual embeddings, word alignments and machine translation quality estimation methods. We test our system by participating in all classification tasks and all language pairs.

The rest of this paper is organized as follows. Section 2 presents detailed information on our translation memory cleaning system. The results of experimental analysis showing the performance of our system is given in Section 3 . The official results of the shared task and the ranking of our system can be seen in Section 4 .

2. System Description

In this section, we present the system used for our submission to the TM cleaning task. Initially, the input data is pre-processed as described in Section 2.1.. The features related to the semantics of the translation pair are extracted by composing the cross-lingual sentence embedding corresponding to each sentence (i.e. source and its translation) and comparing these by several distance metrics, as described in Section 2.2.1.. The second set of features are obtained by TMop considering information about word alignments between the source and target element of each TU, as discussed in Section 2.2.2.. All features are combined and used to predict the category label of each TU by using an ensemble decision tree classifier (Section 2.3.).

2.1. External training resources

We constructed three parallel corpora for the three language combinations considered in the NLP4TM shared task. These are: English-Spanish, English-German and English-Italian.

The English-Spanish and English-Italian corpora were formed using Europarl v.7 (Koehn, 2005), EMEA (Tiedemann, 2009), JRC-Acquis (Steinberger et al., 2006), News Commentary v.11 (WMT 2015) (Bojar et al., 2015) and TED Talks (Cettolo et al., 2012).

The English-German corpus was formed using KDE4, GNOME, OpenOffice, PHP, Ubuntu, Tatoeba (Tiedemann, 2012), Europarl v.7 (Koehn, 2005), CommonCrawl (WMT 2013) (Bojar et al., 2013), News Commentary v.11 (WMT 2015) (Bojar et al., 2015), MultiUN (Eisele and Chen, 2010), DCEP (Hajlaoui et al., 2014), DGT-TM (Steinberger et al., 2013), ECDC-TM (Steinberger et al., 2014) and EAC-TM (by EU Directorate General for Education and Culture).

The external data resources were combined with the task-specific training data provided by the organizers of NLP4TM 2016 (Barbu et al., 2016). All data used to train and test our system are tokenized and lowercased before the feature extraction step using the text processing system of Moses (Koehn et al., 2007). The resulting corpora for each language (given in Table 2.1.) were used to train word embeddings (Section 2.2.1.) and word alignments (Section 2.2.2.).

Language	#Parallel sentences	#Tokens
English-German	11,302,026	469,124,414
English-Spanish	7,632,182	412,341,443
English-Italian	7,967,798	417,857,598

Table 1: Corpus statistics

2.2. Features

2.2.1. Bilingual Embedding Features

The bilingual word embeddings are obtained using the extension of Skipgram model by Luong et al. (2015). The embeddings are trained using the default parameters described by the authors, with a dimension of 200 and using the word alignment option, where the alignments are obtained by fast-align (Dyer et al., 2013). Training of the word embeddings and the word alignments were accomplished using the external data resources described in Section 2.1..

During the computation of a sentence embedding, each word's embedding forming the sentence is averaged after removal of stop words and punctuation in the sentence. The embeddings belonging to the two sentences (one in English and one in the target language) are compared using vector space distance metrics and features of these vectors. The 16 semantic features implemented using these bilingual embeddings make use of:

- Cosine distance, Euclidean distance, Manhattan distance, Chebyshev distance, Canberra distance, number of elements in the sentence vectors, mean of the sentence vectors, median of the sentence vectors.

2.2.2. TMop Features

The second part of the translation memory cleaning system focuses on the features computed by TMop (Jalili et al., 2016) [1]. These features are mainly derived by word alignments (C. de Souza et al., 2013), and quality estimation features (Mehdad et al., 2012; C. de Souza et al., 2014; Turchi et al., 2014; C. de Souza et al., 2015). There are 26 features produced by TMop filters as listed below.

- Sentence length ratio, reverse sentence length ratio, word ratio, reverse word ratio, tag finder output, repeated characters, repeated words, language identifier output, aligned proportion, bi-gram aligned proportion, number of unaligned sequences , longest aligned sequence, longest unaligned sequence, aligned sequence length, unaligned sequence length, first unaligned word, last unaligned word

2.3. Classification

In order to assign the correct class label of translation memory entries, our system uses an Extremely Randomized Trees (ET) classifier (Geurts et al., 2006) as the supervised learning method. The ET classifier applies bagging on the training data and fits a number of decision trees on different subsets to produce the output class as an ensemble average over the decision trees. To build the ET classification model we use 43 features obtained by the Church-Gale score (as computed by one of the baselines proposed by the task organizers), the bilingual embedding features and the TMop features. As optimization metric in our classification model we use the F_1 score.

3. Experiments

The performance of our translation memory cleaning system is evaluated by 10-fold cross-validation on the training set. Each experiment is performed by varying one of the classifier settings and the number of features. We repeat each experiment among all language pairs and classification tasks.

We select three baseline systems to compare the performance of our system. Baseline 1 and 2 are provided by the organizers of the shared task, where the first one uses random label assignment and the second corrects the scores of baseline 1 according to the Church-Gale score (Gale and Church, 1993) computed for each sentence. The 3rd baseline uses majority-voting (i.e. assigning class 1 to all TUs) to predict the label of each translation pair. The classification accuracy of all baseline systems computed with 10-fold cross-validation on the training set can be seen in Table 2.

3.1. Effect of individual feature sets

In the first set of experiments, the features to be used in classification are optimized on the English-Spanish training data set. The systems are evaluated individually with cross-validation by using a different feature (i.e. the semantic features obtained by the distance between two sentence embeddings, and the features obtained by TMop), enabling us to observe the effect of each feature set and the improved performance through their combination.

The results of this analysis are presented in Table 2. Experiments highlight the strong performance of word alignment

[1] Open-source software available at https://github.com/hlt-mt/TMOP

System	F_1 Score		
	Task 1	**Task 2**	**Task 3**
Baseline 1	0.7143	0.7897	0.5250
Baseline 2	0.7296	0.8074	0.5554
Baseline 3	0.8103	0.8724	0.5519

Features	# Features	F_1 Score		
		Task 1	**Task 2**	**Task 3**
Semantic features	16	0.8086	0.8767	0.7637
TMop features	26	0.8549	0.9139	0.8041
All features	43	0.8569	0.9227	0.8233

Table 2: Experiment Set 1: On the left, performance of baseline systems in English-Spanish translation pairs. On the right, F_1 score computed with system predictions and true labels using different sets of features on English-Spanish data. Task 1: binary classification-I, Task 2: binary classification-II, Task 3: fine-grained classification.

EN-DE

System	F_1 Score		
	Task 1	**Task 2**	**Task 3**
Baseline 1	0.7930	0.8603	0.6415
Baseline 2	0.7921	0.8593	0.6436
Baseline 3	0.8751	0.9187	0.6808
w/o FS, w/o SW	0.8754	0.9200	0.8550
w/ FS, w/o SW	-	0.9187	0.8588
w/o FS, w/ SW	**[0.8774]**	**[0.9197]**	**[0.8600]**
w/ FS, w/ SW	-	0.9187	0.8444

EN-ES

System	F_1 Score		
	Task 1	**Task 2**	**Task 3**
Baseline 1	0.7144	0.7898	0.5250
Baseline 2	0.7296	0.8075	0.5555
Baseline 3	0.8103	0.8724	0.5519
w/o FS, w/o SW	0.8536	**[0.9227]**	0.8233
w/ FS, w/o SW	0.8613	0.9164	0.8219
w/o FS, w/ SW	0.8595	0.9226	0.8248
w/ FS, w/ SW	**[0.8621]**	0.9159	**[0.8258]**

EN-IT

System	F_1 Score		
	Task 1	**Task 2**	**Task 3**
Baseline 1	0.6165	0.8278	0.4554
Baseline 2	0.6275	0.8447	0.4852
Baseline 3	0.7642	0.8880	0.4726
w/o FS, w/o SW	0.8164	0.9313	0.7649
w/ FS, w/o SW	0.8080	0.9316	0.7619
w/o FS, w/ SW	**[0.8164]**	**[0.9320]**	**[0.7681]**
w/ FS, w/ SW	0.8081	0.9285	0.7572

Table 3: Experiment Set 2: Performance of the system in English-German, English-Spanish and English-Italian language pairs and all tasks using all features (see Section 2.3.), with the effect of feature selection and sample weighting illustrated. Task 1: binary classification-I, Task 2: binary classification-II, Task 3: fine-grained classification.

features, achieving higher accuracy without the semantic features; although the improved performance when they are combined in the overall system proves the complementary effect of each set of features.

3.2. Effect of feature selection and sample weighting

After determining the features, the possible improvement by adding feature selection to the classifier is tested for the English-Spanish, English-German and English-Italian data sets. The feature selection is implemented by Randomized Lasso (Meinshausen and Bühlmann, 2010) with 10-fold cross-validation and 1000 re-samples. Moreover, due to a highly balanced distribution of labels in the training data, we apply sample weighting on the training data and analyze its effects on the performance. The experiment results showing the feature selection and sample weighting effects are illustrated in Table 3.

The results show that sample weighting increases the performance of the classifier in 8 of the 9 experiments. Sample weighting proves to be a useful choice in the design of the system for this task where the training set consist of highly unbalanced class distributions. On the other hand, feature selection does not provide a significant improvement in 8 of the 9 experiments. This is due to the fact that the ET classifier already has an internal mechanism to achieve the same function, and hence, removes the necessity to have an additional layer of feature selection.

We see that using the cross-validation experiments over the training set our system is able to perform better than all baselines set in each task and its performance slightly varies with changed classifier settings. The best system to participate in each task and language pair is selected as the one with the highest F_1 score among all systems (shown with bold letters in Table 3). The selected systems are trained with the task training data and the label for each TU is predicted on the test data.

4. Results

The official results of our system's participation in each task is given in Table 3.2.. As the official evaluation metric was not announced during the organization of the shared task, we optimized our classifier using F_1 score calculated with respect to the positive class. Therefore, we present the accuracies according to the positive class and also add the weighted F_1 scores, which was selected as the ranking metric by the organizers of the shared task after the submission of the results. Our system achieves the best F_1 scores in the 2nd task with an homogeneous performance among the language pairs. In the light of this result, our system offers a promising application in cleaning of translation memories where the correct label is accepted as little to no post-editing effort, which is a useful perspective to adopt in developing CAT tools.

Our team also manages to obtain the best ranking among all teams in the 1st task with English-Spanish data. The worst performance is still above 0.87 with English-German data.

Task 1				
EN-DE				
System	**Precision**	**Recall**	F_1	F_1 **avg.**
Baseline 1	0.78	0.78	0.78	0.51
Baseline 2	0.78	0.78	0.78	0.51
FBK HLT-MT	**0.78**	**0.98**	**0.87**	**0.49**
EN-ES				
System	**Precision**	**Recall**	F_1	F_1 **avg.**
Baseline 1	0.65	0.66	0.65	0.45
Baseline 2	0.68	0.66	0.67	0.50
FBK HLT-MT	**0.87**	**0.96**	**0.91**	**0.77**
EN-IT				
System	**Precision**	**Recall**	F_1	F_1 **avg.**
Baseline 1	0.63	0.64	0.63	0.52
Baseline 2	0.65	0.64	0.65	0.55
FBK HLT-MT	**0.70**	**0.97**	**0.82**	**0.66**

Task 2				
EN-DE				
System	**Precision**	**Recall**	F_1	F_1 **avg.**
Baseline 1	0.86	0.87	0.86	0.53
Baseline 2	0.86	0.86	0.86	0.53
FBK HLT-MT	**0.85**	**0.99**	**0.92**	**0.49**
EN-ES				
System	**Precision**	**Recall**	F_1	F_1 **avg.**
Baseline 1	0.74	0.75	0.74	0.43
Baseline 2	0.78	0.75	0.76	0.51
FBK HLT-MT	**0.87**	**0.96**	**0.91**	**0.77**
EN-IT				
System	**Precision**	**Recall**	F_1	F_1 **avg.**
Baseline 1	0.80	0.78	0.79	0.50
Baseline 2	0.82	0.78	0.80	0.56
FBK HLT-MT	**0.89**	**0.98**	**0.94**	**0.80**

Task 3			
EN-DE			
System	**Precision**	**Recall**	F_1 **avg.**
Baseline 1	0.64	0.64	0.64
Baseline 2	0.63	0.63	0.63
FBK HLT-MT	**0.68**	**0.77**	**0.70**
EN-ES			
System	**Precision**	**Recall**	F_1 **avg.**
Baseline 1	0.48	0.48	0.48
Baseline 2	0.53	0.52	0.52
FBK HLT-MT	**0.73**	**0.76**	**0.72**
EN-IT			
System	**Precision**	**Recall**	F_1 **avg.**
Baseline 1	0.47	0.47	0.47
Baseline 2	0.50	0.50	0.50
FBK HLT-MT	**0.67**	**0.71**	**0.66**

Table 4: Official Results of the 1^{st} translation memory cleaning shared task of NLP4TM 2016. Two tables on top: binary classification accuracies (precision, recall and F_1) for the positive class and the weighted F_1. Bottom table: Weighted accuracies at the fine-grained classification task. Task 1: binary classification-I, Task 2: binary classification-II, Task 3: fine-grained classification.

The low performance with English-German processing is mainly due to the choice of different domain data in training of the system. This indicates the importance of choosing the data with correct domain when developing our system.

In the fine-grained classification task, we obtain homogeneous results which provide at least 32% relative improvement above the baseline. According to the positive class measurement, in all tasks and language pairs we manage to be beat all baselines. According to weighted F_1, our system provides a relative improvement of 20 - 54% in task 1, 43 - 51 % in task 2 and 9 - 38 % in task 3.

5. Conclusion

We have presented our system that participated in the 1st translation memory cleaning shared task of NLP4TM 2016. Our system exploits a combination of features based on bilingual word embeddings, quality estimation features and word alignments. These features are used to build an ensemble decision tree classifier to capture the semantic and syntactic similarities between the source and target element of a translation unit. We have analyzed our model through two sets of experiments where we vary the classifier and feature settings to measure the correct classification accuracies. The performance of our system in the shared task provides the most stable results in the binary classification-II category, which suggests that this is the most suitable application for the developed system. We show the importance of choosing the domain of training corpus similar to the test data and improvement in performance by using sample weighting. Our team has achieved the 1st ranking in the binary classification-I for English-Spanish category.

6. References

Barbu, E., Escartín, C. P., Bentivogli, L., Negri, M., Turchi, M., Federico, M., Mastrostefano, L., and Orasan, C. (2016). 1st Shared Task on Automatic Translation Memory Cleaning. In *Proceedings of the 2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, Portorož, Slovenia, May.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, 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.

C. de Souza, J. G., Esplà-Gomis, M., Turchi, M., and Negri, M. (2013). Exploiting Qualitative Information from Automatic Word Alignment for Cross-lingual NLP Tasks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics-Short Papers (ACL Short Papers 2013)*, pages 771–776.

C. de Souza, J. G., González-Rubio, J., Buck, C., Turchi, M., and Negri, M. (2014). FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, Maryland, USA, June.

C. de Souza, J. G., Negri, M., Ricci, E., and Turchi, M. (2015). Online Multitask Learning for Machine Translation Quality Estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–228, Beijing, China, July.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. Association for Computational Linguistics.

Eisele, A. and Chen, Y. (2010). Multiun: A multilingual corpus from united nation documents. In Daniel Tapias, et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). Dcep-digital corpus of the european parliament. In *LREC*, pages 3164–3171.

Jalili, M. S., Negri, M., Turchi, M., and Barbu, E. (2016). An unsupervised method for automatic translation memory cleaning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (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*, pages 177–180. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. pages 79–86. MT summit.

Luong, T., Pham, H., and Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Mehdad, Y., Negri, M., and Federico, M. (2012). Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*, pages 171–180, Montréal, Canada, June.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2013). Dgt-tm: A freely available translation memory in 22 languages. *arXiv preprint arXiv:1309.5226*.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., and Gilbro, S. (2014). An overview of the european unionâs highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.

Tiedemann, J. (2009). News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.

Turchi, M., Anastasopoulos, A., C. de Souza, J. G., and Negri, M. (2014). Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Baltimore, Maryland, USA, June.