1st Shared Task on Automatic Translation Memory Cleaning Preparation and Lessons Learned

Eduard Barbu[1], Carla Parra Escartín[2], Luisa Bentivogli[3],
Matteo Negri[3], Marco Turchi[3], Marcello Federico[3],
Luca Mastrostefano[1], Constantin Orasan[4]

[1]Translated, Italy; [2]Hermes Traducciones, Spain; [3]FBK Trento, Italy;
[4]University of Wolverhampton, United Kingdom

{eduard,luca}@translated.net, carla.parra@hermestrans.com,
{bentivo,negri,turchi,federico}@fbk.eu, C.Orasan@wlv.ac.uk

Abstract

This paper summarizes the work done to prepare the first shared task on automatic translation memory cleaning. This shared task aims at finding automatic ways of cleaning TMs that, for some reason, have not been properly curated and include wrong translations. Participants in this task are required to take pairs of source and target segments from TMs and decide whether they are right translations. For this first task three language pairs have been prepared: English → Spanish, English → Italian, and English → German. In this paper, we report on how the shared task was prepared and explain the process of data selection and data annotation, the building of the training and test sets and the implemented baselines for automatic classifiers comparison.

Keywords: Translation Memories, data selection, data annotation

1. Introduction

Translation Memories (TMs) are among the most used tools by professional translators, if not the most used. The underlying idea of TMs is that a translator should benefit as much as possible from previous translations by being able to retrieve how a similar sentence was translated before. Moreover, the usage of TMs aims at guaranteeing that new translations follow the client's specified style and terminology. However, in order to ensure that professional translators can benefit from the contents already stored in a TM, this must be properly maintained and clean.

The first edition of the Natural Language Processing for Translation Memories (NLP4TM 2015) workshop organized at RANLP 2015 (Orasan and Gupta, 2015) highlighted the need for automatic methods for cleaning TMs. For this reason, in the second edition of the NLP4TM workshop (NLP4TM 2016)[1] a shared task on cleaning translation memories has been organized in an attempt to make the creation of resources for TMs easier as well as to enhance TM curation. This paper summarizes how the data for the shared task has been created and how the shared task has been organized.

The remainder of this paper is organized as follows: Section 2. summarizes the shared task. Section 3. shows how we have selected the data (Subsection 3.1.) to be annotated for three language pairs English-Italian, English-Spanish and English-German. The Subsections 3.2. and 3.3. discuss the annotation of the data and the inter-annotator agreement respectively. Section 4. shows how we have made the training and test sets, Section 5. reports on the baselines we have established to measure the participants' system submissions. The final section 6. summarizes our preparatory work for the shared task.

[1] http://rgcl.wlv.ac.uk/nlp4tm2016/

2. Shared Task

The NLP4TM 2016 shared task on cleaning translation memories aims at finding automatic ways of cleaning TMs that for some reason have not been properly curated and include wrong translations. Participants in this task are required to take pairs of source and target segments from TMs and decide whether they are right translations. For this first task three language pairs have been prepared: English → Spanish, English → Italian, and English → German.

The data was annotated with information on whether the source and target content of each TM segment represent a valid translation. In particular, the following 3 point scale has been applied:

1. The translation is correct (tag "1").
2. The translation is correct, but there are a few orthotypographic mistakes and therefore some minor post-editing is required (tag "2").
3. The translation is not correct (content missing/added, wrong meaning, etc.) (tag "3").

For each language pair, two thirds of the annotated segments are provided for training and one third is provided for testing during the evaluation phase.

Besides choosing the pair of languages with which they want to work, participants can choose to participate in either one or all of the following three tasks:

1. **Binary Classification (I)**: In this task, it is only required to determine whether a segment is right or wrong. For the first binary classification option, only tag ("1") is considered correct because the translators do not need to make any modification, whilst tags ("2") and ("3") are considered wrong translations.

2. **Binary Classification (II)**: As in the first task, in this task it is only required to determine whether the segment is right or wrong. However, in contrast to the first task, a segment is considered correct if it was labeled by annotators as ("1") or ("2"). Segments labeled ("3") are considered wrong because they require major post-editing.
3. **Fine-grained Classification**: In this task, the participating teams have to classify the segments according to the annotation provided in the training data: correct translations ("1"), correct translations with a few orthotypographic errors ("2"), and wrong ("3").

Participants were required to register their intention to participate by filling in an online form. Upon registration, we provided the registered participants with the training set. The test set will be distributed during the evaluation phase and the participating teams will be asked to submit the output of their systems in a format similar to the training set[2]. For evaluation, the standard measures precision, recall and the F1 will be used. In addition, we have foreseen a potential manual error analysis of subsets of the test data. The extent of this analysis will depend on the number of systems submitted. The numbers of runs submitted by participants has not been limited, although the participating teams are required to indicate their primary (and secondary, if relevant) runs.
In order to ensure the reusability and replicability of the shared task results and with the aim of making a real impact in professional translation workflows, all participants have been encouraged to release their systems and make them publicly available for future use. Besides, the development of methods that can be run on large datasets without requiring a lot of computational resources is also fostered. Thus, participans have also been encouraged not to use machine translation as one of the factors used to determine the class of a segment.

3. Data preparation

3.1. Data selection

The data was sampled from the public part of MyMemory (Trombetti, 2009) the biggest translation memory database in the world. The public part of MyMemory is composed of all bi-segments that the translators agreed to make public, from public parallel corpora and glossaries, data crawled from parallel sites on the web and the individual contributions through a collaborative web interface.
Regarding the percentage of errors, the bi-segments coming from the translators have fewer errors, the bi-segments coming from the collaborative web interface have most errors and the bi-segments coming from public parallel corpora or from crawling the web are somewhere in the middle.
In the initial phase we extracted approximately 30K bi-segments for each language pair taking care to sample from all the above mentioned sources. The bi-segments are heterogeneous and belong to different domains ranging from medicine and physics to colloquial conversations. Once we had this first pre-selection, we filtered the extracted bi-segments according to the following criteria:

1. **Minimum length**. The source and target segments should contain at least three words. MyMemory contains a significant number of entries that have only a word or two. However in many cases it is hard to understand if the source is a translation of the target because the context for interpreting the source and target is missing. We decided to avoid this situation for the task and therefore all segments shorter than a 3-word-span were deleted.

2. **No tags**. The extracted bi-segments should not contain tags or strange characters. Even if in the translation memory cleaning task one should consider segments that contain tags or strange characters, their identification is trivial and therefore was excluded from the task.

3. **Appropriate language codes**. The language codes of the source and target segments should coincide with the declared language codes. For example, if the source segment language code is declared as English and the target language code segment is declared as Spanish then the source segment language code should be English and the target segment language code should be Spanish. To check that this is indeed the case we used the high quality automatic language detector Cybozu [3].

4. **One to Many/Many to One**. We only accepted those bi-segments where one source sentence corresponds to at least one target sentence or one target sentence corresponds to at least one source sentence. That is: all bi-segments where many sentences in the source segment corresponded to many sentences in the target sentence were rejected because these bi-segments need realignment.

5. **Uniqueness**. The source and target segments should be unique across the set. We allowed the possibility of having a repeated source segment with multiple corresponding target segments as long as the target segments differed from each other, and viceversa: a unique target segment with diferring source segments[4].

From the bi-segments that met the above criteria we sampled again 10K bi-segments per language pair from which we then manually selected approximately 3K bi-segments per language pair. To facilitate the manual selection of the negative examples, we computed the cosine similarity score between the Machine Translation of the English segment and the target bi-segment. The hypothesis to consider was that low cosine similarity scores can signal bad translations.

[2] Due to time constraints, the testing phase will take place in the last weeks prior to the NLP4TM 2016 workshop and therefore no results can be reported at this time.

[3] https://github.com/shuyo/language-detection

[4] Two segments are different if the segments as character string are different after space normalization.

The manually selected bi-segments do not contain inappropriate language or other errors that cannot be identified automatically.

3.2. Data annotation

The set containing approximately 3K bi-segments per language pair was annotated by two native speakers of each target language. The guidelines for annotating this data set contain annotation instructions and examples[5]. In what follows, we present the annotation guidelines for English–Spanish. Similar annotation guidelines have been produced for the English–German and English–Italian language pairs.

1. You should give the score "1" if the translations can be accepted without editing. That is, the segment in Spanish preserves the meaning of the English segment.

   Example: "This product contains mineral oil."→"Este producto contiene aceite mineral." is a good Spanish translation of the English original segment. You do not need to change anything: punctuation or words.

2. You should give the score "2" when the few operations of editing you perform do not affect the meaning of the phrase. For example you should annotate "2" when:

   - The Spanish segment preserves the meaning of the English segment. However the Spanish segment has very few extra stuff that once deleted makes the translation acceptable:

     Example: "This product contains mineral oil."→"d Este producto contiene aceite mineral.". Deleting the "d" at the beginning makes the translation acceptable (tag "1")

   - The Spanish segment has (or lacks) punctuation that however do not impede understanding the segment. Adding or deleting the extra-punctuation renders the translation acceptable (tag "1"):

     Example: "This product contains mineral oil."→"Este producto contiene aceite mineral". Adding the final dot renders the translation (tag "1").

   - The Spanish segment has very few typos relative to the length of translation. Correcting the typos makes the translation acceptable (tag "1").

     Example: "This product contains mineral oil."→"Este produto contiene aceite mineral.". Correcting produto→producto makes the translation acceptable (tag "1").

3. You should give the score "3" if you need to perform substantial editing or editing that changes the meaning of the Spanish segment.

[5]The reader can consult these annotation guidelines at the web address: http://rgcl.wlv.ac.uk/nlp4tm2016/shared-task/.

Annotator	Annotator 2			
	Category	1	2	3
Annotator 1	1	**1127**	276	281
	2	209	**382**	305
	3	10	9	**360**

Table 1: The agreement for English–Italian

   - *Example*: "This product contains mineral oil."→"Este producto contiene agua mineral.". You need to replace a whole content word that is "agua" (water) with a new word "aceite" (oil) and thus the meaning of the sentence changes.

   - *Example*: "This product contains mineral oil."→"Este produto contiene aceite mineral". In this case, you need to change produto→producto, aciete→aceite and add the final dot to render an acceptable translation. Even if the editing operations do not change the meaning of the Spanish segment the numbers of edits you need to perform is substantial relative to the length of the segment.

The annotation has been performed with the aid of the MT-Equal (Girardi et al., 2014), a toolkit for Human Assessment of Machine Translation Output, developed and maintained by FBK. MT-Equal is an online tool accessible through the Chrome web browser[6]. It defines two types of users: administrators and annotators. While the annotators perform the annotation, the administrators can load data, assign tasks to the annotators, follow the task progress, export the results etc.
Our initial idea was that after the two annotators annotate the 3K they will agree on more than 2K bi-segments. The identical annotated bi-segments would then be used to build the training and test sets. In the next section, we discuss the inter-annotator agreement for each language-pair.

3.3. Inter-annotator agreement

We computed the inter-annotator agreement using the well known Cohen's kappa coefficient (Cohen, 1960). In Table 3.3., we present the agreement for the English–Italian language pair. The main diagonal of the table shows the number of bi-segments where the annotators[7] agree. They agreed for 1869 bi-segments. The number fell short of the 2K bi-segments we were expecting. To reach at least that number we asked an arbiter to annotate the 281 bi-segments that were annotated with tag 1 by annotator 1, and with tag 3 by annotator 2. The arbiter annotated 182 instances with 1, 32 instances as 2 and 67 instances as 3. The final set to be used for training and testing for English–Italian consists of the sum of all agreements and the arbiter resolution (2118 bi-segments). The Cohen's kappa coefficient for the English–Italian annotation task is 0.41.
The initial English–Spanish set had 3012 bi-segments. The first annotator annotated all bi-segments whereas the second annotator annotated 2708 bi-segments. The annotator

[6]http://mtequal.fbk.eu/

[7]labeled Annotator 1 and Annotator 2, respectively

Annotator	Annotator 2			
	Category	1	2	3
Annotator 1	1	**1413**	63	166
	2	203	**193**	107
	3	64	29	**470**

Table 2: The agreement for English–Spanish

Annotator	Annotator 2			
	Category	1	2	3
Annotator 1	1	**1629**	131	13
	2	23	**42**	3
	3	3	10	**15**

Table 3: The agreement for English–German

agreement is calculated for the 2708 common annotations and is reported in table 3.3..

The set to be used for training and testing for English-Spanish consists of the sum of all agreements (2076 bi-segments). The Cohen's kappa coefficient for the English–Spanish annotation task is higher than the same coefficient for English–Italian 0.57.

The initial English–German set had 3016 bi-segments. The first annotator annotated 2509 bi-segments and the second annotator annotated only 2404 bi-segments. However, the annotators chose to work on a different order and while annotator 1 started from the first segment, the second annotator chose to perform the annotation in reverse order, starting by the last bi-segment. The annotator agreement is calculated for the 1869 common annotations in table 3.3..

The Cohen's kappa coefficient for the English–German annotation task is 0.37. Two things can be observed relative to Table 3.3.: the number of bi-segments for which we have agreement is less than 2K (1686), just like in the English–Italian case, and the number of negative bi-segments (annotated with 3 by both annotators) is very low (15). To have a training and test sets comparable with the training and test sets for the other language pairs (English–Italian and English–Spanish), we added noise to the English–German set. We took 410 bi-segments annotated by one of the annotators and not by the other and added noise such as to transform them in 300 bi-segments annotated with 3 and 109 bi-segments annotated with 2. The set to be used for training and testing for English–German consists of the set of all bi-segments where both annotators agreed plus the 410 bi-segments to which we added noise (2096 bi-segments in total).

In conclusion, we selected three sets containing approximately 2K bi-segments where two annotators agreed. According to the interpretation that Landis and Koch (Landis and Koch, 1977) give to Cohen' kappa coefficient, the reported agreement coefficients is borderline between poor and fair. We have not conducted a study to see why the agreement is low. However inspecting a sample of disagreement cases we have noted that the annotators disagree when the translators bring into the translation process background knowledge that is not stated explicitly in the source sentence. For example the word "drug" in the source language can be translated as "the drug for dogs" in the target language when the information that the drug was meant to be for dogs was stated in the context before the segment to be translated.

4. Training and Test Sets

The training and test have been built using stratified sampling. This means that the training and test sets contain the same percentage of bi-segments with the same category label. Table 4. gives the number of bi-segments having the category labels "1", "2" and "3" in the training and test sets for all language pairs. The names of the columns E–G, E–S and E–I stand for English–German (E–G), English–Spanish (E–S) and English–Italian (E–I), respectively.

	Language Pair			Category Label
Training Set	E–I	E–S	E–G	
	872	942	1086	1
	254	128	100	2
	284	313	210	3
Test Set	E-I	E-S	E-G	
	437	471	544	1
	128	65	51	2
	143	157	105	3

Table 4: The size of the training and test sets

5. Baseline systems

To benchmark the results of the classifiers that the participants to the Shared Task will submit we have implemented two baselines. The first baseline generates random labels for the test set with the same distribution of the labels in the training set. The second baseline corrects the results of the first baseline when the Church-Gale (Gale and Church, 1993) score of the source and target segments is above a predefined threshold fixed to 2.5[8]. The idea is that if the difference in length between the source and target segments is too big, then it is likely that the target segment is not the translation of the source. Therefore in these cases we modified the score given by the first baseline to "3". To measure the length of the source and destination segments, we use the modified Church-Gale length difference algorithm (Tiedemann, 2011) presented in Equation 1:

$$CG = \frac{l_s - l_d}{\sqrt{3.4(l_s + l_d)}} \quad (1)$$

The results of the two baselines for all the shared tasks defined in section 2. are presented in table 5..

As stated earlier, we compute Precision, Recall and the F1 score for the two baselines defined before and each sub task defined in the shared task. It is expected that baseline 2 is harder to beat than baseline 1 (baseline 2 gains at most 3 points of F-score over baseline 1). With the exception of the Fine-Grained task, the baselines are not easy to beat, as they reach, in the case of the Binary Classification approximately 0.8 F-score points.

[8]The script that computes the baselines can be downloaded from the URL http://rgcl.wlv.ac.uk/resources/NLP4TM2016/baselines.py.remove

	Language Pair			Measure
	E–I	E–S	E–G	
Baseline 1 Fine-Grained	0.45	0.52	0.63	P
	0.45	0.52	0.63	R
	0.45	0.52	0.63	F1
	E-I	E-S	E-G	
Baseline 2 Fine -Grained	0.47	0.55	0.62	P
	0.47	0.55	0.62	R
	0.47	0.55	0.62	F1
	E-I	E-S	E-G	
Baseline 1 Binary Classification 1	0.61	0.68	0.78	P
	0.62	0.69	0.78	R
	0.61	0.69	0.78	F1
	E-I	E-S	E-G	
Baseline 2 Binary Classification 1	0.62	0.71	0.78	P
	0.62	0.69	0.77	R
	0.62	0.70	0.78	F1
	E-I	E-S	E-G	
Baseline 1 Binary Classification 2	0.8	0.77	0.85	P
	0.79	0.77	0.86	R
	0.79	0.77	0.85	F1
	E-I	E-S	E-G	
Baseline 2 Binary Classification 2	0.82	0.80	0.85	P
	0.79	0.77	0.85	R
	0.80	0.78	0.85	F1

Table 5: Baselines for the shared task

6. Conclusion

In this paper we have presented the methodology for constructing three sets of parallel bi-segments for English–Italian, English–Spanish and English–German sampled from the MyMemory translation memory database. We expected a higher agreement in the annotation task but due to time constraints to release the data for the shared task, we could not assess properly why the level of disagreement was so high. For English–Italian we needed an arbiter to decide a number a cases and thus achieve around 2K annotated examples where at least two annotators agreed in their annotations. The English–German set did not contain enough negative examples , meaning that MyMemory has good quality segments for this language pair. We have created some artificial negative segments by adding noise to the acceptable ones. We have implemented and presented two baselines to be compared against the classification results sent by the participants in the shared task.

7. Acknowledgments

The research reported in this paper is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n° 317471. Part of the work has been supported by the EC-funded project ModernMT (H2020 grant agreement no. 645487). We are grateful to Translated for giving us access to the MyMemory database. Last but not least we want to thank the 6 annotators who have annotated the data.

8. Bibliographical References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *COMPUTATIONAL LINGUISTICS*.

Girardi, C., Bentivogli, L., Farajian, M. A., and Federico, M. (2014). Mt-equal: a toolkit for human assessment of machine translation output. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, August 23-29, 2014, Dublin, Ireland*, pages 120–123.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Constantin Orasan et al., editors. (2015). *Proceedings of the First Workshop on Natural Language Processing for Translation Memories (NLP4TM-2015)*, RANLP 2015, Hissar, Bulgaria, September.

Tiedemann, J. (2011). *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, USA.

Trombetti, M. (2009). Creating the world's largest translation memory.