The first Automatic Translation Memory Cleaning Shared Task

Noname manuscript No.
(will be inserted by the editor)

The First Automatic Translation Memory Cleaning Shared Task

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

Received: date / Accepted: date

Abstract This paper reports on the organization and results of the first Automatic Translation Memory Cleaning Shared Task. This shared task is aimed at finding automatic ways of cleaning translation memories (TMs) that have not been properly curated and thus include incorrect translations. As a follow up of the shared task, we also conducted two surveys, one targeting the teams participating in the shared task, and the other one targeting professional translators. While the researchers-oriented survey aimed at gathering information about the opinion of participants on the shared task, the translators-oriented survey aimed to better understand what constitutes a good TM unit and inform decisions that will be taken in future editions of the task. In this paper, we report on the process of data preparation and the evaluation of the automatic systems submitted, as well as on the results of the collected surveys.

Keywords Translation Memories · Translation Memory Cleaning · Natural Language Processing

[1]
Translated, Italy
E-mail: eduard@translated.net

[2]
ADAPT Centre, SALIS/CTTS, Dublin City University, Ireland
E-mail: carla.parra@adaptcentre.ie

[3]
FBK Trento, Italy
E-mail: {bentivo,negri,turchi,federico}@fbk.eu

[4]
University of Wolverhampton, United Kingdom
E-mail: C.Orasan@wlv.ac.uk

1 Introduction

Translation memories (TMs) are databases that store previously translated segments. They are usually integrated in computer-assisted translation (CAT) tools and professional translators use them to retrieve past translations of the same or similar segments. They are among the most used repositories of information by professional translators. An example of a TM is MyMemory [30], the largest TM in the world. While as of March 2016 the DGT TM [28] had 111,225,619 translation units (TUs), as of September 2016, MyMemory has over 1,483,925,012 TUs stored and it is used by 10 million users monthly. At least 10,000 of these users are professional translators, and these users perform around 50 million searches.

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 kept clean. Yet, manual cleaning of the translation memories is costly and for specialized TMs it requires an expertise that is hard to find. Translators shall not only be acquainted with the terminology and style of a particular client, but they also have to be able to detect mistranslations and inconsistencies in an efficient and timely manner. Normally, such tasks are carried out by the proofreaders, i.e. the translators in charge of correcting the translations of their fellow colleagues, as they are more used to spotting problems and correcting them.

However, this task still constitutes a bottleneck and moreover, the fast pace at which TMs grow nowadays makes it also not feasible to perform manual TM cleaning processes. Just as an example, MyMemory receives approximately 15 million contributions per month. Given the size of modern TMs and the impossibility of performing efficient manual cleaning tasks, an automatic solution is therefore necessary.

The purpose of the first Automatic Translation Memory Cleaning Shared Task was to invite teams from both academia and industry to tackle this problem and submit their automatic systems for evaluation. As this was the first shared task on this topic, this year we also focused on learning to define the task better and on understanding what are the most promising approaches to tackle the problem. The proposed task consisted in identifying the translation units that had to be discarded because they were not accurate translations of each other, or corrected as they contained orthotypographical errors like missing punctuation marks or misspellings.

Examples 1a–1b and 2a–2c taken from the English-Spanish part of MyMemory, illustrate this. As indicated below, Examples 1a and 1b show TUs that should be deleted. This is because in both cases, the original English sentence was mistranslated. In the case of 1a, the Spanish translation of the English source sentence says "Date of the last update of this summary:", while in 1b there is an orthographic mistake ("dias" instead of "días"), and the Spanish translation says "I hope that you have good days". Examples 2a–2c, on the other hand, require minor orthotypographical corrections as can be observed when comparing the existing target sentence of the TU and the corrected one below in each case. Example 2a is missing the opening question mark required in Spanish, and the accent on the interrogative *cómo*, Example 2b only requires the deletion of the extra character "e" appearing at the beginning of the target sentence, and in Example 2c the word *Español* (Spanish) needs to be lowercased.

1. TUs to be discarded
   (a) **Example 1**.
   EN: This summary was last updated on 11 February 2009.
   ES in TM: Fecha de la última actualización del presente resumen:
   ES (right): El presente resumen se actualizó por última vez el 11 de febrero de 2009.
   (b) **Example 2**.
   EN: have a good night
   ES in TM: que pases buenos dias
   ES (right): que pases una buena noche / que duermas bien
2. TUs to be corrected
   (a) **Example 3**.
   EN: Hello how are you doing
   ES in TM: hola como te va
   Es (right): hola, ¿cómo te va?
   (b) **Example 4**.
   EN: 6.1 List of excipients
   ES in TM: e 6.1 Lista de excipientes
   ES (right): 6.1 Lista de excipientes
   (c) **Example 5**.
   EN: My Spanish teacher is called
   ES in TM: mi profesor de Español se llama
   ES (right): mi profesor de español se llama

TUs containing minor syntactic mistakes have been studied in the past and some people working in industry have developed tools to identify them. Two well known quality assurance (QA) and terminology tools[1] are ApSIC Xbench[2] and Verifika[3]. These pricey tools are designed to support the manual identification and correction of errors in TMs and ongoing translations. More concretely, these tools implement a series of formatting and terminology checks for all segments in a TM or in an ongoing translation and allow the user to

[1] Most Computer Assisted Translation tools like SDL Trados Studio (http://www.sdl.com/cxc/language/translation-productivity/trados-studio/) and memoQ (https://www.memoq.com/) also include specific QA modules.

[2] http://www.xbench.net/

[3] https://e-verifika.com/

amend such errors. They check, for example, if an opened tag has its corresponding closing tag, if a word is repeated, or if a word is misspelled. Their users can also use terminological glossaries as a reference and the tools will then verify that the terms included in the glossary are translated accurately. Finally, users can manually define their own QA rules by means of regular expressions. These may check, for instance, that Spanish adjectives agree with their accompanying nouns in gender and number.

However, to the best of our knowledge none of these tools address the semantic problems of the segments that are not proper translations of each other. The task we propose also bears similarity with two other tasks defined in the literature: machine translation quality estimation (MTQE) [11,4, 27, among others] and bilingual document alignment [9]. Both tasks are part of the Workshop on Machine Translation that is organized yearly. While the MTQE shared task was already established in 2012 [8], the bilingual document alignment is a new shared task proposed in 2016 [5].[4] The main difference between our task and these other two is that we deal with translations produced by professional translators and not machines,[5] and that TMs, despite containing entire documents, are more heterogeneous and the TUs do not always include information about the context in which each segment appears. Although we acknowledge that some of the techniques and features used in the above-mentioned tasks can be successfully reused for automatic TM cleaning tasks, the task has therefore some features that make it unique.

The remainder of this paper is structured as follows: Sections 2 and 3 summarize the shared task and the data used. Section 4 focuses on the evaluation metrics employed for the shared task and the established baselines. Section 5 offers an overview of the participating teams and the strategies they used, while section 6 discusses the results obtained. Section 7 reports on the surveys we did as a follow up of the shared task, and finally Section 8 summarizes and wraps up our work.

2 Task description

The NLP4TM 2016 Automatic Translation Memory Cleaning Shared Task aimed at finding automatic ways of cleaning TMs that for some reason have not been properly curated and thus include incorrect translations.

For this first task, TUs for three frequently used language pairs were prepared: English → Spanish; English → Italian; and English → German.

The data was annotated with information on whether the target content of each TU represents a valid translation of its corresponding source. In particular, the following 3-point scale was applied:

[4] `http://www.statmt.org/wmt16/bilingual-task.html`

[5] Although in some cases machine translation may have been used to produce translations, translators have to verify that such translations are correct before they are stored as new TUs in a TM.

1. The translation is correct (tag “1”).[6]
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 and should be discarded (content missing/added, wrong meaning, etc.) (tag “3”).

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

1. **Binary Classification (I)**: In this task, it was only required to determine whether a TU was correct or incorrect. For this binary classification option, only tag (“1”) was considered correct because the translators do not need to make any modification, whilst tags (“2”) and (“3”) were considered incorrect translations.
2. **Binary Classification (II)**: As in the first task, in this task it was only required to determine whether the TU was correct or incorrect. However, in contrast to the first task, a TU was considered correct if it was labeled by annotators as (“1”) or (“2”). TUs labeled (“3”) were considered incorrect because they require major post-editing.
3. **Fine-grained Classification**: In this task, the participating teams had to classify the TUs according to the annotation provided in the training data: correct translations (“1”), correct translations with a few orthotypographic errors (“2”), and incorrect translations (“3”).

In order to ensure the re-usability and replicability of the shared task results and with the aim of making a real impact in professional translation workflows, all participants were encouraged to release their systems and make them publicly available for future use.[7] The development of methods that can be run on large datasets without requiring a lot of computational resources was also fostered. Thus, participants were also encouraged not to use machine translation as one of the factors used to determine the class of a TU. However, and as discussed later in Section 5, one team tried the computationally-intensive neural MT approach.

3 Data

The data was in the most part sampled from the public part of MyMemory, the biggest translation memory database in the world. The public part of MyMemory is composed of all TUs that the translators agreed to make public,[8]

[6] In the absence of sufficient context, any translation which had some context in which it would be adequate was accepted.

[7] Unfortunately, to our knowledge as of September 2016 only one of the participating teams has released their system (the JUMT Team). The FBK system was trained using the open source TM cleaner TMop [13]. All systems are described in the working notes available at http://rgcl.wlv.ac.uk/nlp4tm2016/working-notes-on-cleaning-of-translation-memories-shared-task/.

[8] See: https://mymemory.translated.net/doc/en/tos.php

public parallel corpora and glossaries, data crawled from parallel sites on the web and the individual contributions through a collaborative web interface.

With regards to the percentage of errors contained in MyMemory, it was noticed that the TUs coming from the translators have the fewest errors, the TUs coming from the collaborative web interface have the most errors, and the TUs coming from public parallel corpora or from crawling the web are somewhere in the middle.

In the initial phase we extracted approximately 30,000 TUs for each language pair taking care to sample from all the above mentioned sources. The TUs were heterogeneous and belonged to different domains ranging from medicine and physics to colloquial conversations. Once we had this first pre-selection, we automatically selected a subset of TUs 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. The choice of the minimum word-span length was arbitrary and thus further research is needed to confirm if it should be expanded.
2. **No tags**. The extracted TUs should not contain tags or strange characters. Even though in a translation memory cleaning task one should consider segments that contain tags or strange characters, their identification is trivial and therefore we decided not to address this in our task.
3. **Appropriate language codes**. The actual language 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.[9] Cybozu was selected because in our experience it gives better results than other language detectors including Google's Compact Language Detector.[10]
4. **One to Many/Many to One**. We selected only those TUs where one source sentence corresponded to at least one target sentence or one target sentence corresponded to at least one source sentence. This is the case when the source sentence is too long and the target language requires the splitting of its translation into two sentences (i.e. *one-to-many*), or the other way round: two short source-language sentences are joined into a longer target-language sentence (i.e. *many-to-one*). On the other hand, all TUs where many sentences in the source segment corresponded to many

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

[10] For a benchmark see `http://blog.mikemccandless.com/2011/10/accuracy-and-performance-of-googles.html`.

sentences in the target (i.e. *many-to-many*) were rejected because these TUs needed 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 differing source segments.[11]

From the TUs that met the above criteria we sampled again 10,000 TUs per language pair from which we then manually selected approximately 3,000 TUs per language pair. Since the proportion of TUs containing incorrect translations is low, to facilitate their manual selection we computed the cosine similarity score between the machine translation of the English segment and the target segment of the TU. The hypothesis to consider was that low cosine similarity scores (less that 0.3) can signal bad translations.

Finally, we ensured that the manually selected TUs did not contain inappropriate language or other errors that could not be identified automatically.

3.1 Data annotation

The set containing approximately 3,000 TUs per language pair was annotated by two native speakers of each target language. The guidelines for annotating each data set contained annotation instructions and examples.[12] Although we are aware of the existence of fine-grained and detailed frameworks for classifying translation errors such as the TAUS Dynamic Quality Framework (DQF), the Multidimensional Quality Metrics (MQM) [7] and their harmonized version, recently released [16], we did not take them into consideration because the overall aim was not to identify error types, but rather identify TUs that needed to be discarded. For future editions, this may however be an interesting issue to explore and take into account.

The annotation was performed with MT-EQuAl [12], a toolkit for human assessment of machine translation output developed and maintained by FBK. MT-EQuAl is an online tool accessible through a web browser.[13] 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 annotated the 3,000 TUs they would agree on more than 2,000. The identically annotated TUs would then have been used to build the training and test sets. However, our hypothesis only held true for the English → Spanish language pair. For English → Italian we had an arbiter that annotated the TUs where the initial

[11] Two segments are different if the segments as character strings are different after space normalization.

[12] The annotation guidelines are available at: `http://rgcl.wlv.ac.uk/nlp4tm2016/shared-task/`

[13] `http://mtequal.fbk.eu/`

raters disagreed and for English → German we encountered that the majority of the segments in which both annotators agreed belonged to the "correct" category (label "1"), which left us with too few TUs in the data belonging to the categories semi-correct ("2") and wrong ("3"). To compensate for this lack of negative data, we had a native speaker that added extra noise in the data.[14]

The data annotation required the equivalent to approximately one full week of labour per annotator. When it was required to carry out additional tasks such as re-annotating part of the data with a new annotator (i.e. in the case of English → Italian), or adding noise to create segments falling into the underrepresented categories (i.e. in the case of English → German), additional time was required. For future editions of the shared-task we would like to increase the size of the data. A bigger data sample would also allow us to assess whether the current training/testing data is enough.

3.2 Training and Test Sets

For each language pair, two thirds of the annotated TUs were provided for training and one third was provided for testing during the evaluation phase.

The training and test sets were built using stratified sampling. This means that the training and test sets contain the same percentage of TUs with the same category label. Table 1 gives the number of TUs having the category labels "1", "2" and "3" in the training and test sets for all language pairs. The names of the columns, EN → DE, EN → ES and EN → IT stand for English → German, English → Spanish and English → Italian, respectively.

	Language Pair			Category Label
Training Set	EN → IT	EN → ES	EN → DE	
	872 (62%)	942 (68%)	1086 (78%)	1
	254 (18%)	128 (9%)	100 (7%)	2
	284 (20%)	313 (23%)	210 (15%)	3
	1410	*1380*	*1396*	
Test Set	EN→IT	EN→ES	EN→DE	
	437 (62%)	471 (68%)	544 (78%)	1
	128 (18%)	65 (9%)	51 (7%)	2
	143 (20%)	157 (23%)	105 (15%)	3
	708	*693*	*700*	

Table 1 Size of the training and test sets. The percentage of each category with respect to the total is indicated in parenthesis.

[14] The inter-annotator agreement results and a more detailed report of how the data was prepared can be found in [3].

4 Evaluation metrics and baselines

As shown in Table 1, the distribution of positive and negative examples in the task data is highly unbalanced with the majority of points assigned to the positive class (*"1"*). Please notice that this setting overestimates the distribution of the negative examples in the MyMemory database for these language pairs. However if we had tried to reflect the percentage of positive and negative examples in the database it would have resulted in test and training sets with very few negative examples thus making it impossible to properly evaluate the classifiers.

In this condition, classic evaluation metrics used for classification, such as F_1 score and accuracy, tend to result in high uninformative scores that prevent a reliable evaluation of the submitted systems. In order to cope with this issue, we hence opted for two task-oriented evaluation metrics, averaged F_1 score and balanced accuracy (computed as the sum of recalls for each class divided by the number of classes), which give equal weight to each class.

To better analyse the results of the participating systems, we implemented two baselines. The first baseline generates random category labels (*i.e.* 1, 2, or 3) 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 [10] score of the source and target segments is above a predefined threshold fixed to 2.5.[15] The idea is that if the normalized 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 the TU was marked as incorrect (by assigning the category label 3).

To measure the length of the source and destination segments, we used the modified Church-Gale length difference algorithm [29] presented in Equation 1:

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

5 Participants

As shown in Table 2, six teams participated in at least one of the proposed sub-tasks, by submitting a total of 45 runs. All teams participated in one sub-task (EN → DE Binary Classification II) and three of them (FBK, Lingua Custodia and Jadavpur/Saarland University) participated in all sub-tasks. Table 3 provides a classification of the approaches adopted by participants in terms of the type of approach, the data and the resources used.

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

	EN → DE			EN → IT			EN → ES		
	BTI	BTII	FGT	BTI	BTII	FGT	BTI	BTII	FGT
Autodesk	-	1	1	-	1	1	-	1	1
Univ. of Edinburgh	2	2	2	-	-	-	-	-	-
FBK HLT-MT	1	1	1	1	1	1	1	1	1
Jadavpur/Saarland Univ.	1	1	-	1	1	-	1	1	-
Lingua Custodia	1	1	1	1	1	1	1	1	1
Univ. of South Africa	1	1	1	1	1	1	1	1	1

Table 2 Automatic Translation Memory Cleaning Shared Task participants and number of submitted runs, divided by language pair and sub-task. Due to space limitations, the task names have been abbreviated here to *BTI* (binary task I), *BTII* (binary task II), and *FTG* (fine-grained task).

	Autodesk	**Univ. of Edinburgh**	**FBK HLT-MT**	**Jadavpur/ Saarland Univ.**	**Lingua Custodia**	**Univ. of South Africa**
Paradigm	ML	ML	ML	Rules	ML	Rules + ML
Algorithm	Random forest	Random forest	Extremely Randomized Trees	-	Random forest	SVM
Quality indicators	Segment length, surface mismatches, MT features (character, word, PoS overlap between translated source and the target segment)	Surface features, language model, word alignment, neural MT	Segment length, word length, surface mismatches, sentence similarity, word alignment	Segment length, surface mismatches	Segment length, word length, surface mismatches, sentence similarity, word alignment	Segment length, surface mismatches, grammaticality, LM probabilities, lexical translation probabilities
Tools	PoS tagger, in-house Moses-based MT system	Neural MT, word aligner, language model, subword unit splitter	Open source TM cleaner, word aligner	-	Word aligner, stemmer	Spellchecker, quality assurance tool, grammar checker
Resources	task data, in-house data	task data, external bilingual, and monolingual data	task data, external bilingual data, external TMs	task data	task data, dictionaries	task data, external bilingual data

Table 3 Overview of the different systems submitted classified by team. *ML* stands for machine learning, *POS* for part of speech, and *LM* for language model.

In the remainder of this Section we present more details about the participating systems:

Autodesk [32] submitted a system based on a machine learning approach that extends the features proposed in [2] and uses them in a random forest classifier.[16] The new features leverage automatic translations of the source segment in the target language to reward translation units with a substantial overlap between the automatic translation and the target segment. The overlap is computed at character, word and Part-of-Speech (PoS) level using the machine translation evaluation metric BLEU [20] and the Levenshtein distance. The machine translation system is an in-house system based on Moses trained on Autodesk data. PoS information is obtained using the TreeTagger [23] and the Universal Tagset proposed by [22].

The **University of Edinburgh** (Uedin) [6] participated with an approach that leverages several groups of features previously developed in machine translation quality estimation. These features measure source and target complexity and misalignments between source and target segments taking advantage of neural components, such as word alignment and machine translation. Similarly to other submissions, these features are then used in a random forest classifier. To limit the impact of rare words in source and target languages, the training and test sentences are pre-processed by splitting each word in subword units (*i.e.* sequence of characters) as proposed by [24]. They submitted two systems that make use of the same features but differ on the MT system used: one uses a statistical MT system, while the other uses a neural MT model.

FBK HLT-MT [1] proposed a learning-based approach based on extremely randomized trees as the classification algorithm.The submitted system is based on the TMop[17] open source TM cleaning tool, which integrates three main groups of feature extractors. The first group captures translation quality by checking the correctness of the source and target language and by looking at surface aspects, such as the possible mismatches in the number of numbers, dates, URLs and XML tags present in the two segments. The second group focuses on translation fluency and includes features based on word alignment information to link source and target words and capture the quantity of meaning preserved by the translation [26]. The third group focuses on translation adequacy by considering the similarity of bilingual sentence embeddings. By using the method proposed in [25], cross-lingual word embeddings return a common vector representation for words in different languages. This representation is used in different ways to measure the similarity of the source and target segments (e.g. by computing cosine similarity). Sample weighting is also applied to cope with the unbalanced distribution of training data.

[16] The random forest classifier, similar to the extremely randomized trees, is an ensemble learning method that minimises overfitting by combining the output of multiple decision trees in a single class label. The two algorithms slightly differ in the way they split the trees (in a deterministic way in the case of random forest and randomly in the case of extremely randomized trees).

[17] https://github.com/hlt-mt/TMOP

The **Universities of Jadavpur and Saarland** (JUMT Team) [19], for their joint participation, adopted a rule-based approach in which information about source and target lengths, word lengths and capitalisation / punctuation / number mismatches are explicitly modelled instead of being supplied as input features for a classifier. Their system has been publicly released in GitHub.[18]

Lingua Custodia [17] participated with a machine learning solution based on random forest as the classification algorithm. The features used consider source and target lengths, token lengths, mismatches at the level of punctuation, numbers and capitalised letters, spacing issues, word similarity at character level (after stemming with Snowball[19]), sentence similarity and alignment scores (computed with Hunalign[20]). The adopted learning method also takes advantage of instance weighting to cope with the unbalanced data distribution and feature selection to avoid overfitting (performed with the Random Feature Elimination algorithm available in Scikit-Learn [21]). An interesting outcome of the applied feature selection process is that Church-Gale scores, word stems similarity and alignment scores are the most predictive ones across tasks.

The **University of South Africa** (Unisa) [31] participated with a machine learning-based approach that combines a variety of features to train the Scikit-Learn implementation of Super Vector Mashine (SVM) classifiers. Some of the features look at the length of the source and target (e.g. Gale-Church ratio and normalised length difference). Others are obtained from the rule-based translation quality checking methods implemented in the **pofilter** tool[21] and from rule-based spelling (Hunspell[22]) and grammar (LanguageTool[23]) checkers. Statistical features using external data to identify fluency issues are also included. They range from normalised language model probabilities to lexical translation probabilities learned from the Europarl corpus [14].

6 Results and Discussion

The results for all tasks included in the shared task (binary task I, binary task II and fine-grained task) are presented in tables 4, 5 and 6. The tables show Averaged F_1 and balanced accuracy scores for all participating teams and the two baselines.[24]

[18] `https://github.com/nayakt/TMCleaning`

[19] http://www.nltk.org/_modules/nltk/stem/snowball.html

[20] https://github.com/danielvarga/hunalign

[21] http://docs.translatehouse.org/projects/translate-toolkit/en/stable-1.14.0/commands/pofilter.html

[22] https://hunspell.github.io/

[23] https://languagetool.org/

[24] For consistency between the binary tasks and the fine-grained task, the Averaged F_1 score for the fine grained task corresponds to macro-averaged F_1 score. The macro average computes the average precision, recall or F_1 score whereas the micro average first sums the true positives, true negatives, false positives and false negatives and only then computes the average precision, recall or F_1 score. Due to space restrictions we cannot present all the measures computed to evaluate the performance of each system. For a more detailed

In all the tasks, most of the submitted systems outperform the baselines providing a significant improvement both in averaged F_1 and balanced accuracy. To test if these differences in performance are by chance, we compared the proportion of the negative examples between each classifier and the most competitive baseline using the McNemar's test [18,15]. The Null Hypothesis we want to reject at a significance level of 0.001 is that there is no difference between the balanced accuracies of each submitted system and the most competitive baseline. We successfully rejected the null hypothesis in all cases except for 4 cases of the 6 submissions by the JUMT Team. Since the JUMT Team system follows a rule-based approach, this might be due to the fact that the rules designed by manually analysing the training/development set were not sufficient to cover the phenomena present in the test set.

Nevertheless with the exception of one submission (the English → German binary classification task I), the P-value computed by the test is close to the highly significant level.

Considering the three language pairs, English → German has shown the smallest improvements with respect to the baselines and in particular on the binary classification tasks I and II. This may be due to the fact that *i)* the percentage of negative examples in the training and test data sets is smaller than in the other language pairs (15% for English → German, 21% for English → Italian, and 22% for English → Spanish), and that *ii)* German is a more complex language compared to English, Italian and Spanish because it is highly inflected, it has a different word order and it makes use of compound words. All these factors can affect the quality of the extracted features and the capability of the learning algorithms of correctly classifying the negative test samples.

Looking at the results of the submitted systems, there is not one participant that performs the best in all tasks and language pairs. This suggests that each language pair has its own peculiarities and errors that require *ad hoc* solutions. Analysing the behaviour of each submission, we notice that the use of hand-crafted rules proposed by the JUMT Team results in the largest variance in terms of performance confirming the difficulty of porting rules between languages. Although taking advantage of machine translation was discouraged in the call for participants, when used it has shown a clear impact on the performance bringing such systems among the top ranked. However, there is a clear difference between the best performing system and the least accurate system for each submission. The difference ranges from 6 to 18 points of the accuracy score. This fact, if combined with the observation that we do not have a clear winner, invites the conclusion that features of different systems can be combined and that a new system that implements these features will perform better in all the tasks.

presentation, the interested reader can consult the overview summary available at `http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/results-1st-shared_task.pdf`

TEAM	Averaged F_1			Balanced Accuracy		
	EN → IT	EN → ES	EN → DE	EN → IT	EN → ES	EN → DE
Baseline1	0.51	0.43	0.50	0.51	0.45	0.50
Baseline2	0.54	0.50	0.23	0.54	0.43	0.23
FBK HLT-MT	0.65	0.76	0.49	0.65	0.64	0.52
JUMT Team	**0.75**	0.72	0.62	**0.75**	0.73	0.65
Lingua Custodia	0.71	**0.81**	0.63	0.71	**0.8**	0.61
Unisa	0.74	0.75	**0.71**	0.73	0.73	**0.67**
Uedin1	-	-	0.69	-	-	0.66

Table 4 Results for the binary classification task I.

TEAM	Averaged F_1			Balanced Accuracy		
	EN → IT	EN → ES	EN → DE	EN → IT	EN → ES	EN → DE
Baseline1	0.49	0.43	0.52	0.49	0.43	0.52
Baseline2	0.55	0.50	0.52	0.56	0.5	0.52
Autodesk	**0.84**	**0.80**	0.47	**0.85**	**0.76**	0.5
FBK HLT-MT	0.8	0.76	0.48	0.75	0.73	0.51
JUMT Team	0.69	0.65	0.57	0.77	0.7	0.63
Lingua Custodia	0.82	0.78	0.64	0.8	0.75	0.6
Unisa	0.76	0.76	**0.68**	0.71	0.72	**0.64**
Uedin1	-	-	0.66	-	-	0.63

Table 5 Results for the binary classification task II.

TEAM	Averaged F_1			Balanced Accuracy		
	EN → IT	EN → ES	EN → DE	EN → IT	EN → ES	EN → DE
Baseline 1	0.35	0.29	0.32	0.35	0.28	0.32
Baseline 2	0.39	0.34	0.32	0.39	0.34	0.32
Autodesk	0.57	0.52	0.32	0.57	0.51	0.34
FBK HLT-MT	0.55	0.52	0.34	0.53	0.50	0.36
JUMT Team	-	-	-	-	-	-
Lingua Custodia	**0.63**	**0.66**	0.46	**0.62**	**0.64**	0.43
Unisa	**0.63**	0.6	**0.58**	0.59	0.56	**0.52**
Uedin1	-	-	0.49	-	-	0.46
Uedin2	-	-	0.48	-	-	0.46

Table 6 Results for the fine-grained classification task.

7 Post-hoc survey and a look at the future

With the aim of gathering further information on the shared task itself and on what constitutes a good TU, we conducted two separate surveys, one targeting the teams participating in the shared task, and one targeting professional translators.

7.1 The translators-oriented survey

The translators-oriented survey aimed at better understanding what constitutes a good TU for those who work with TMs on a daily basis. The survey consisted of a grid specifically targeting the TU quality which asked the respondent to mark what was or was not a correct, a wrong and an almost correct TU and the additional four following questions:

1. What is your background?
2. When working with TMs, do you usually encounter TUs that you think should be corrected and/or deleted (e.g. bad translations, misalignments, etc.)?
3. Do you actively curate the TMs you work with to ensure they are of a high quality?
4. What is the most common mistake/error you see in stored TUs?

We gathered a total of 309 replies. 274 respondents (88.7%) indicated that they were either translators, proofreaders or revisers. 14 (4.5%) were translation industry specialists or researchers, 7 (2.3%) were researchers in academia, 2 (0.6%) researchers in industry and 12 (3.9%) indicated that they were none of the above and reported to be engineers, project managers, social entrepreneurs, lawyers and interpreters or translators and researchers.

291 respondents (94.2%) indicated that in their daily work they usually encountered TUs that should be corrected or deleted, while only 18 (5.8%) indicated that they did not. As to whether they performed TM curation tasks, 250 (80.9%) replied that they did, while 59 (19.1%) indicated that they did not.

The most interesting results come from the "free text" question asking for the most common mistake they saw stored in TUs. We manually processed all replies and grouped them together in major categories. It shall also be indicated that some replies had to be discarded due to a clear reference to machine translation output rather than a TM (e.g. "*MT for German–Romanian is basically unusable*"). Moreover, some respondents referred to more than one error and therefore the number of replies to this question outnumbers the total number of respondents to the survey. We considered a total of 345 errors, which can be divided into linguistic (307, 89%) and non-linguistic (38, 11%) errors.

Figure 1 illustrates the distribution of the types of errors identified by the respondents. We classified all errors across 12 main categories. It shall be noted that the same TU may have several types of errors at once. We refer here to the errors mentioned by the respondents as the most common mistake they encountered in TUs. The typology was drafted when analysing the replies and thus differs from other standard proposals such as the DQF and MQM referenced to in Section 3.1. Moreover, as it was a free text question and we did not focus on translation errors solely but *any* type of error that could be encountered when revising a TU, some error categories not included in these

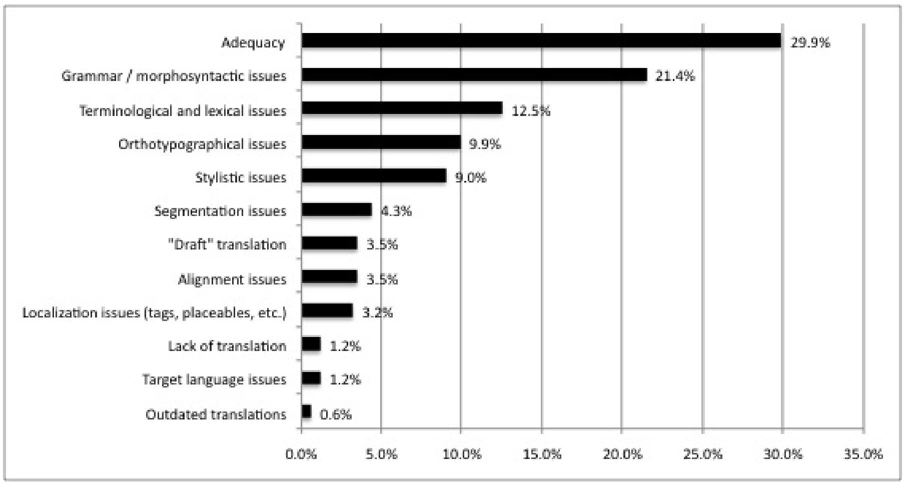

Fig. 1 Main errors found in TMs according to professional translators.

typologies were detected (alignment issues, segmentation issues, localisation issues, outdated translations).

1. “Draft” translation
2. Adequacy
3. Alignment issues
4. Grammar / morphosyntactic issues
5. Lack of translation
6. Localization issues (tags, placeables, etc.)
7. Orthotypographical issues
8. Outdated translations (old translations that need to be updated to comply with newer glossaries/conventions/style guides, etc.)
9. Segmentation issues
10. Stylistic issues
11. Target language issues
12. Terminological and lexical issues

As may be observed, the most frequent error relates to adequacy issues (29.9%). Replies under this category included complaints about bad and inconsistent translations, mistranslations and incoherent translations or translations not taking into account the domain of the source text. The second biggest category relates to grammar and morphosyntactic issues (21.4%), and the third one to terminological and lexical issues (12.5%). The fourth and fifth most common errors related to orthotypographical issues (9.9%) and stylistic issues (9%) respectively. It seems therefore clear, that for professional translators the most important issues to be tackled when cleaning translation memories are indeed of a linguistic nature and more specifically, issues related to the meaning of the segments and their grammatical correctness.

As mentioned earlier, the translators were also asked to fill in a grid about the quality of TUs. All statements in the grid started with "A translation unit where the translation element..." and for each ending of the sentence, the translators had to mark whether that statement corresponded to a correct, a wrong or an almost correct TU. The following 30 endings were used:

1. ... is syntactically perfect.
2. ... has minor adequacy issues.
3. ... has minor fluency issues.
4. ... does not require editing.
5. ... requires the correction of at most 1 word.
6. ... requires the correction of at most 10% of the words (e.g. 2 out of 20).
7. ... requires the correction of at most 15% of the words (e.g. 3 out of 20).
8. ... requires the correction of at most 20% of the words (e.g. 4 out of 20).
9. ... requires the correction of at most 50% of the words (e.g. 10 out of 20).
10. ... contains punctuation errors.
11. ... contains casing errors.
12. ... contains extra/missing/untranslated words.
13. ... contains at most one missing/untranslated words.
14. ... contains at most 10% missing/untranslated words (e.g. 2 out of 20).
15. ... contains at most 15% missing/untranslated words (e.g. 3 out of 20).
16. ... contains at most 20% missing/untranslated words (e.g. 4 out of 20).
17. ... contains at most 50% missing/untranslated words (e.g. 10 out of 20).
18. ... contains at most 10% extra words (e.g. 2 out of 20).
19. ... contains at most 15% extra words (e.g. 3 out of 20).
20. ... contains at most 20% extra words (e.g. 4 out of 20).
21. ... contains at most 50% extra words (e.g. 10 out of 20).
22. ... contains redundant sentences.
23. ... contains extra/missing negations, changing the sentence polarity.
24. ... can be influenced by the document context resulting in a translation that disambiguates necessary parts of the segment to ensure cohesion/coherence in the target language (e.g. the source segment contains pronouns but they are replaced by the referent in the translation: "It has 4 legs" → "The dog has 4 legs", where it is clear from the document context that the pronoun "it" refers to the noun "dog").
25. ... contains more than a sentence.
26. ... contains one or both segments tokenized (i.e. extra blank spaces separate words and punctuation marks/characters).
27. ... contains irrelevant material that can be deleted in one go (for example extra words added at the end of the sentence).
28. ... contains spelling mistakes that can be easily fixed using a spellchecker.
29. ... contains obviously wrong numbers and dates that can be fixed with a few keystrokes.
30. ... is semantically equivalent to the source segment.

Figure 2 shows the replies of the translators. As can be observed, there is a lot of disagreement as to what is a correct or a wrong segment, although three statements clearly stand out: “is syntactically perfect”, “does not require editing”, and “is semantically equivalent to the source segment”. Similarly, it seems that four statements clearly identify wrong TUs: “requires the correction of at most 50% of the words (e.g. 10 out of 20)”, “contains at most 50% missing/untranslated words (e.g. 10 out of 20)”, “contains at most 50% extra words (e.g. 10 out of 20)”, and “contains extra/missing negations, changing the sentence polarity”.

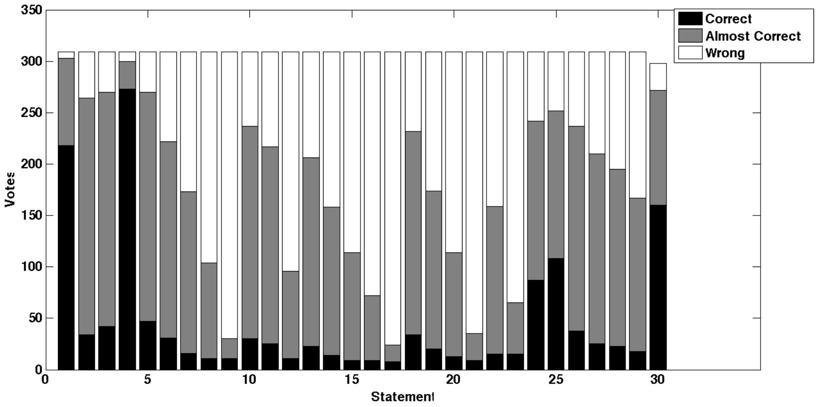

Fig. 2 What is the quality of TUs? – Translators grid.

7.2 The researchers-oriented survey

The researchers-oriented survey aimed at gathering general information about the participating teams and their systems, but also on the organization of the shared task itself. Similarly to the translators, the teams were asked to fill in a grid on their concept of quality of a TU. The same statements as with the translators were used. However, in this case only two options were given: “yes/true”, for those cases in which the statement held true, and “no/false”, for those in which it did not. Figure 3 shows the results. As can be observed, all teams agreed that correct TUs “are syntactically perfect” and “require the correction of at most 1 word” and all wrong TUs “contain punctuation errors”, “contain at most 10% extra words (e.g. 2 out of 20).”, and “contain redundant sentences”. Again, for the rest of the cases, there is a lot of disagreement and in one case one team left one of the questions unanswered (23). In the only case in which both, translators and the participating teams in the shared task seem to agree, is that correct TUs should be syntactically perfect.

All but one shared task participants agreed that the shared task was useful and the one who did not indicated that it may be useful. As regards to whether or not the distinction between binary and fine-grained classification should be

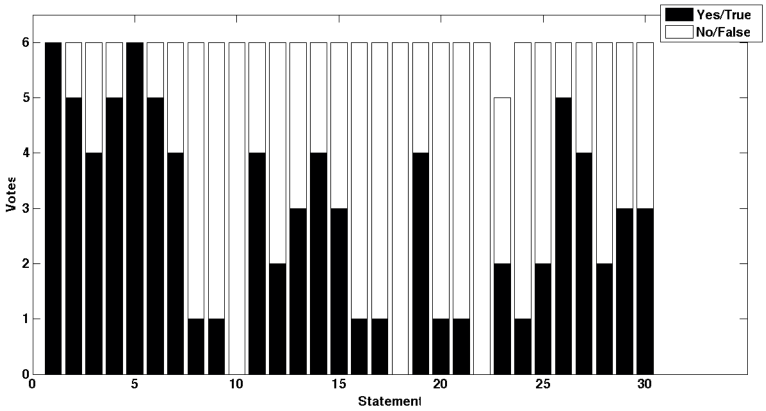

Fig. 3 What is the quality of TUs? – Shared task participants grid.

kept in future editions of the shared task, 3 teams indicated that it should be kept as it is, 2 indicated that it should not, and 1 team was unsure.

The evaluation metrics used for the shared task were also included in the survey. Five teams agreed that the confusion matrix should be used for evaluation and 4 agreed that the weighted F_1 score taking into account the class imbalance should be used. One team suggested to also have the micro F_1 score included, and a different team voted for the macro F_1 score. Finally, one team suggested to use the F_1 score but only on the negative class. For future editions of the shared task this will be taken into account.

From the results of the survey, it seems clear that the participants liked the shared task and would like to see it repeated next year. Some of them suggested to drop the binary classification task II. The fact that several language pairs were included was also positively valued and it was suggested to include further language pairs in future editions. One team also suggested to keep the test set large as this year to have meaningful results.

As to potential improvements, it was suggested to establish a clear evaluation metric from the beginning, as some methodologies tune on that, and also to establish a shorter evaluation period. Some teams also asked for domain-specific TMs. It is true that evaluation metrics should have been defined from the very beginning and that the evaluation period could have been shorter. However, as this was the first time that this shared-task was organized some issues had to be decided along with the shared task itself. In the next edition the evaluation metrics will be defined from the beginning. The long evaluation period was established to allow more teams that had showed interest to participate or had asked for an extension to participate, but ideally it should have been shorter. The use of domain-specific TMs still needs to be explored.

8 Conclusions

In this paper we have reported on the organization of the first Automatic Translation Memory Cleaning Shared Task. Data for three different language pairs (English → Italian, English → German and English → Spanish) was prepared and three tasks were proposed (binary classification I, binary classification II and fine-grained classification).

We have explained how the data was prepared for the task and reported on the results obtained for the 45 runs submitted by the 6 teams participating and we have also reported on the strategies used by the different teams. There was no clear winner in all tasks but there was a marked difference between the best performing system and the least performing one. This invites the conclusion that a better system can be built leveraging the best features of all participating systems. We are currently gathering these features from each participating team.

Two surveys were conducted with the aim of gathering further information about the task. The first survey involved professional translators and aimed at finding the most common errors in TMs. The most important error types accounting for more than 60 percent of errors are related to the translation process, grammar and terminology. In next year's shared task we will try to address these issues by asking the participants to identify the type of error they have found in the bi-segments (e.g. translation error, grammar error in the target segment, inappropriate use of terminology, etc.) and not only annotate the segment as good or bad. The second survey was conducted with the participants to the first shared task. The results were less conclusive, some participants advising to drop the fine grained task but others finding the task interesting and proposing to keep it. To keep the results relevant and to encourage further participation next year we might add new language pairs and annotate more TUs. Furthermore, after the discussion with a translation memory curator, we found out that in some settings it is very important to identify the TUs to be deleted from the translation memory but that it is equally important not to drop too many false negatives. Next year we will use a ranking measure that rewards the systems for the true negative segments found while penalizing them for the false negatives.

Acknowledgements 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^{o} 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. We would also like to thank the anonymous reviewers for their valuable feedback to improve this paper and for the ideas for future work they have provided us with. Last but not least, we want to thank the 6 annotators who annotated the data used in the shared-task.

References

1. Ataman, D., Jalili Sabet, M., Turchi, M., Negri, M.: FBK HLT-MT Participation in the 1st Translation Memory Cleaning Shared Task. Working Notes on Cleaning of Translation Memories Shared Task – http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/fbkhltmt-workingnote.pdf (2016)
2. Barbu, E.: Spotting False Translation Segments in Translation Memories. In: Proceedings of the Workshop Natural Language Processing for Translation Memories, pp. 9–16. Hissar, Bulgaria (2015)
3. Barbu, E., Parra Escartín, C., Bentivogli, L., Negri, M., Turchi, M., Federico, M., Mastrostefano, L., Orasan, C.: 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 (2016)
4. Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., Ueffing, N.: Confidence estimation for machine translation. In: Proceedings of the 20th International Conference on Computational Linguistics (CoLing-2004), pp. 315–321 (2004)
5. Buck, C., Koehn, P.: Findings of the WMT 2016 Bilingual Document Alignment Shared Task. In: Proceedings of the First Conference on Machine Translation, pp. 554–563. Association for Computational Linguistics, Berlin, Germany (2016). URL http://www.aclweb.org/anthology/W/W16/W16-2347
6. Buck, C., Koehn, P.: UEdin participation in the 1st Translation Memory Cleaning Shared Task. Working Notes on Cleaning of Translation Memories Shared Task – http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/ChristianBuck-TM_Cleaning_Shared_Task.pdf (2016)
7. Burchardt, A., Lommel, A.: Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality. Tech. rep., DFKI, Berlin, Germany (2014)
8. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2012 workshop on statistical machine translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 10–51. Association for Computational Linguistics, Montréal, Canada (2012). URL http://www.aclweb.org/anthology/W12-3102
9. Esplà Gomis, M., Forcada, M.L.: Bitextor, a free/open-source software to harvest translation memories from multilingual websites. In: Proceedings of the workshop Beyond Translation Memories: New Tools for Translators MT. Ottawa, Ontario, Canada (2009)
10. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. COMPUTATIONAL LINGUISTICS (1993)
11. Gandrabur, S., Foster, G.: Confidence estimation for translation prediction. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL'03, pp. 95–102. Association for Computational Linguistics, Stroudsburg, PA, USA (2003). DOI 10.3115/1119176.1119189. URL http://dx.doi.org/10.3115/1119176.1119189
12. Girardi, C., Bentivogli, L., Farajian, M.A., Federico, M.: Mt-equal: a toolkit for human assessment of machine translation output. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pp. 120–123. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (2014). URL http://www.aclweb.org/anthology/C14-2026
13. Jalili Sabet, M., Negri, M., Turchi, M., C. de Souza, J.G., Federico, M.: Tmop: a tool for unsupervised translation memory cleaning. In: Proceedings of ACL-2016 System Demonstrations, pp. 49–54. Association for Computational Linguistics, Berlin, Germany (2016). URL http://anthology.aclweb.org/P/P16/P16-4009
14. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit, pp. 79–86. AAMT, AAMT, Phuket, Thailand (2005)
15. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience (2004)
16. Lommel, A.: Multidimensional Quality Metrics (MQM) Definition. Tech. rep., DFKI, Berlin, Germany (2015)

17. Mandorino, V.: The Lingua Custodia Participation in the NLP4TM2016 TM Cleaning Shared Task. Working Notes on Cleaning of Translation Memories Shared Task – http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/description_LinguaCustodia.pdf (2016)
18. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947). DOI 10.1007/BF02295996. URL `http://dx.doi.org/10.1007/BF02295996`
19. Nahata, N., Nayak, T., Pal, S., Naskar, S.: Rule Based Classifier for Translation Memory Cleaning. Working Notes on Cleaning of Translation Memories Shared Task – http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/Working_Note-JUMTTeam.pdf (2016)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311–318. Association for Computational Linguistics (2002)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn : Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
22. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey (2012)
23. Schmid, H.: Treetagger— a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart **43**, 28 (1995)
24. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers (2016)
25. Søgaard, A., Agić, v., Martínez Alonso, H., Plank, B., Bohnet, B., Johannsen, A.: Inverted Indexing for Cross-lingual NLP. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1713–1722. Beijing, China (2015)
26. C. de Souza, J.G., Esplà-Gomis, M., Turchi, M., Negri, M.: 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 (Volume 2: Short Papers), pp. 771–776. Sofia, Bulgaria (2013)
27. Specia, L., Turchi, M., Cancedda, N., Dymetman, M., Cristianini, N.: Estimating the Sentence-Level Quality of Machine Translation Systems. In: 13th Annual Meeting of the European Association for Machine Translation (EAMT-2009), pp. 28–35 (2009)
28. Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlter, P.: DGT-TM: A freely available Translation Memory in 22 languages. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)
29. Tiedemann, J.: Bitext Alignment. No. 14 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, USA (2011). DOI 10.2200/s00367ed1v01y201106hlt014. URL `http://dx.doi.org/10.2200/s00367ed1v01y201106hlt014`
30. Trombetti, M.: Creating the worlds largest translation memory. MT Summit XII – Ottawa (2009)
31. Wolff, F.: Unisa system submission at NLP4TM 2016. Working Notes on Cleaning of Translation Memories Shared Task – http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/UNISA.pdf (2016)
32. Zwahlen, A., Carnal, O., Läubli, S.: Automatic TM Cleaning through MT and POS Tagging: Autodesks Submission to the NLP4TM 2016 Shared Task. Working Notes on Cleaning of Translation Memories Shared Task – http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/nlp4tm-adsk.pdf (2016)