Project funded by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

Project reference: 317471

Project full title: EXPloiting Empirical appRoaches to Translation

D6.3: Improved Corpus-based Approaches

Authors: Carla Parra Escartin (Hermes), Lianet Sepúlveda Torres (Pangeanic)

Contributors: Constantin Orasan (UoW), Anna Zaretskaya (UMA), Santanu Pal (USAAR), Hernani Costa (UMA), Rohit Gupta (UoW), Liling Tan (USAAR), Varvara Logacheva (USFD), Carolina Scarton (USFD), Liangyou Li (DCU), Chris Hokamp (DCU), Joachim Daiber (UvA), Hoang Cuong (UvA), Hanna Bechara (UoW)

***Document Number*:** EXPERT_D6.3_20160921

***Distribution Level*:** Public

***Contractual Date of Delivery*:** 30.06.16

***Actual Date of Delivery*:** 21.09.16

Contributing to the Deliverable: WP6

WP Task Responsible: Hermes

EC Project Officer: Ioanna Peppa

D6.3. Improved corpus-based Approaches

Carla Parra Escartín[1], Lianet Sepúlveda Torres[2]

[1]Hermes Traducciones, Spain; [2]Pangeanic, Spain
carla.parra@hermestrans.com, lisepul@gmail.com

September 21, 2016

Contents

1 Introduction

Corpora constitute one of the key resources for research in Natural Language Processing and Translation Studies alike. As such, they can be the data being investigated, or the data used for carrying out research. In the translation industry, translation memories (i.e. bilingual sentence-aligned corpora) are constantly used for leveraging new translation projects, concordance searches and storing new translations. They are also used as the source data for terminology extraction tools and for training Machine Translation (MT) systems. In the case of Machine Translation, corpora are the key for success. Good curated and large monolingual and bilingual corpora are crucial for training the language and translation models of MT systems respectively.

With the advances in research, new ways of both compiling corpora and using corpora have emerged. Traditional methods focusing on linguistic or heuristic approaches have now evolved towards hybrid methods that aim at combining the best of both worlds. In the EXPERT project, this hybridization has been vastly investigated and new advances and contributions have been made. This Deliverable aims at summarizing the main results as regards to corpus-based approaches of the EXPERT project. Although some overlapping with previous deliverables is unavoidable[1], here we aim at offering an overview of all the advances made across the different Work Packages in the project.

This Deliverable has been organized using the different work packages of the EXPERT project in which work on hybrid corpus-based approaches has been done[2]. Section 2 is devoted to the advances in Work Package 3: *Data Collection*, and Section 3 refers to the ones on Work Package 4: *Language technology, domain ontologies and terminologies*. Section 4 covers Work Package 5, *Learning from and informing translators*, and Section 5, reports on the research carried out within Work Package 6, *Improved corpus-based Approaches*. Finally, Section 6 summarizes the main advances in terms of hybrid corpus-based approaches and suggests new avenues for exploration in this matter.

2 Data collection

As specified in *Annex I - Description of Work* of the EXPERT Project, an additional limitation of data collection approaches is that they "have par-

[1] See `http://expert-itn.eu/?q=content/deliverables`.

[2] Here, we focus on corpus-based approaches exclusively. For an overview of the work done with regard to translation tools, please see Deliverable *D6.4. Improved Hybrid Translation Tool* [Parra Escartín and Sepúlveda Torres, 2016b]. All work involving user evaluations has been summarized in Deliverable *D6.5. Final User Evaluation* [Parra Escartín and Sepúlveda Torres, 2016a].

ticular data constraints which prevent the use of the same data for different approaches". Bearing this in mind, the proposed EXPERT solution was to "investigate how data repositories can be built automatically in a way that makes them useful to multiple corpus-based approaches to translation".

Bilingual and multilingual corpora are crucial for both Translation Studies and research in MT. However, in many cases such corpora are very difficult to find, particularly if the language pair involved is not a widely researched one (e.g. German–Italian), or it involves one or several under-resourced languages. In the case of specialized translation, the problem is also more acute, as the need for bilingual corpora also has the requisite that such corpora are highly specialized and meet certain quality requirements. Thus, it could be the case that large bilingual corpora exist for English–Spanish, but not in the fashion domain, for instance. One strategy to overcome this problem is to automatically compile such corpora on-demand using the so-called comparable corpora (i.e. non-parallel bilingual and multilingual text sources belonging to the same domain), or locating bilingual sources that can subsequently be aligned.

Researchers within the EXPERT project have explored ways of compiling both comparable and parallel corpora. The next two subsections report on the work in this topic. Subsection 2.1 reports on efforts towards compiling comparable corpora and creating tools to do so, and Subsection 2.2 reports on the ones done in the case of parallel corpora.

Finally, an alternative strategy to obtain data for research is to artificially generate it. Subsection 2.3 reports on the different strategies used by EXPERT researchers to produce new, artificial data using already existing resources.

2.1 Comparable Corpora Compilation

ESR3 did a comprehensive state-of-the-art study on comparable corpora compilation in which he not only compared the strategies used in different fields such as Corpus Linguistics, but also the different existing techniques and applications[3]. His ultimate goal was to build a new automatic tool for corpus compilation. In Costa [2015b] and Costa et al. [2015a], ESR3 offers an extensive study of Distributional Similarity Measures (DSMs). While in Costa [2015b] he focuses on the usage of DSMs to describe specialized comparable corpora, in Costa et al. [2015a] he focuses on using them to measure the relatedness between documents in specialized comparable corpora. In both cases, he uses three DSMs: (1) Spearman's Rank Correlation Coefficient (SCC), (2) Chi-Square(X^2), and (3) the Number of Common Entities (NCE). As input for these DSMs, ESR3 used lists of common tokens, lemmas and stems. All experiments reported in Costa [2015b] and Costa et al.

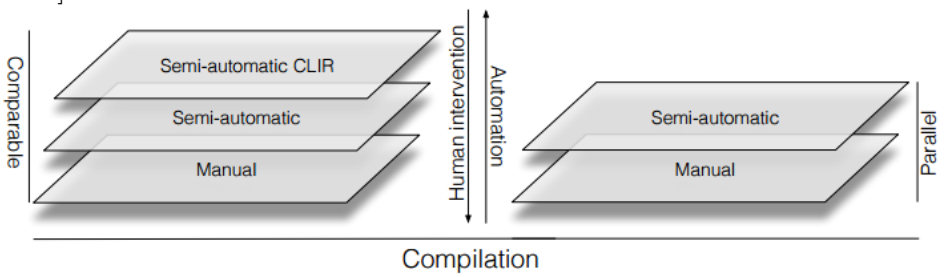
[3]For a more detailed overview on this topic, consult Deliverable *D3.1. Framework for Data Collection* [Costa, 2015a] of the EXPERT project.

[2015a] obtained similar results in terms of performance for all the tested DSMs. As far as the features used as input, the authors conclude that using a list of common tokens is enough to describe the data and that the usage of more processing power and time in integrating tools like stemmers and/or lemmatizers is therefore not needed. Another interesting conclusion in Costa [2015b] is that DSMs can be used as a suitable tool to rank documents by their similarities. In fact, the methodology proposed in Costa [2015b] showed a high performance in the task of filtering out documents with a low level of relatedness.

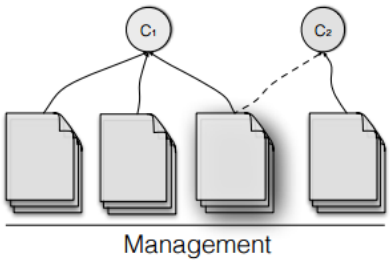
As a follow up of his state-of-the-art work, ESR3 designed and implemented an application to semi-automatically compile mono- and multilingual parallel and comparable corpora, named *iCorpora* [Costa et al., 2015c,b]. *iCorpora* is a Web system that guides the user through the process of compiling, managing and exploring comparable and parallel corpora. *iCorpora* intends to overcome or improve some of the issues found in other corpus compilation tools accessible in the market, like *BootCaT*[4] and *WebBootCat*[5].

As shown in Figure 1, *iCorpora* has three main components: *iCompileCorpora*, *iManageCorpora* and *iExploreCorpora*.

Figure 1: Structure of *iCorpora* and its different components [Costa et al., 2015c]



(a) *iCompileCorpora.*



(b) *iManageCorpora.*

(c) *iExploreCorpora.*

[4] http://bootcat.sslmit.unibo.it/
[5] https://www.sketchengine.co.uk/

The first component, *iCompileCorpora*, is a three-layered model consisting of a manual layer, a semi-automatic web-based layer and a semi-automatic Cross-Language Information Retrieval (CLIR) layer. This design option was chosen with a twofold aim: increasing the flexibility and robustness of the compilation process and hierarchically extending the manual layer features iteratively to the semi-automatic web-based layer and then to the semi-automatic CLIR layer [Costa et al., 2015c]. The *iManageCorpora* component was specially designed to manage corpora, measure the similarity between documents and to explore the representativeness of the corpora. The management activities comprise a set of tasks such as editing documents, copying and pasting sentences and documents across the corpus and dividing corpora into sub-corpora (e.g. according to domains and subdomains) [Costa et al., 2015c]. Finally, the *iExploreCorpora* component offers a set of concordance features, such as searching for words in context and automatically extracting the most frequent words and multiwords [Costa et al., 2015c].

Emerson et al. [2014] investigated a different approach to corpus compilation in which they follow the structural guidelines proposed by Abney and Bird [2011]. The authors built a multilingual corpus, SeedLing, comprising texts in 1451 languages belonging to 105 language families ($\sim$ 19.0% of the world's languages). The authors cleaned data from four different sources: the Online Database of Interlinear Text, the Omniglot website, Wikipedia and the Universal Declaration of Human Rights. To illustrate potential uses of such a corpus, the authors ran an experiment on automatic identification of similar languages with promising results. The resulting corpus as well as their scripts are publicly available in a github repository[6].

2.2 Parallel Corpora Compilation

While comparable corpora may be very useful for some tasks such as terminology extraction and information retrieval, Machine Translation systems require large parallel bilingual corpora to be successfully trained.

ESR2 explored ways of compiling parallel corpora out of comparable ones. In Pal et al. [2014b], the authors explore ways of retrieving parallel corpora from Wikipedia for the language pair English→Bengali. They propose a methodology for automatically aligning parallel text fragments using textual entailment and a Phrase Based SMT (PB-SMT) system. The extracted parallel data is subsequently concatenated to existing training data for a PB-SMT system. In their experiments, this resulted in better overall results in terms of translation quality over the baseline (+1.73 BLEU points, which corresponds to a relative improvement of +15.84%).

The methodology proposed in Pal et al. [2014b] is further developed in

[6] https://github.com/alvations/SeedLing

Pal et al. [2015b], where the authors further investigate the usage of Semantic Textual Entailment and Distributional Semantics for text similarity with the aim of identifying parallel fragments. Their experiments with an English→Bengali SMT system obtained relatively positive results, suggesting that this approach could help to improve the quality of MT systems involving low-resourced language pairs. The authors acknowledge, however, that despite obtaining positive results the evaluation scores were low. One possible explanation for this may be the fact that Bengali is a morphologically rich language with relatively free phrase order. The authors manually inspected a subset of the output produced by their SMT system and confirmed that their approach yields less out-of-vocabulary words and makes better lexical choices than the baseline system.

ER1 has aimed at compiling parallel corpora from the web[7]. To do so, he first did a statistical estimation of the quantity of parallel data available in Common Crawl[8], a large repository of crawled data freely available for research and industrial purposes. He also estimated the precision and recall of three data collection systems broadly used in industry for parallel data collection: the ILSP crawler [Papavassiliou et al., 2013], Bitextor [Esplà-Gomis and Forcada, 2009] and the TAUS data spider [Ruopp and Xia, 2008]. After gathering this information, he implemented a system based on URL matching that automatically identifies parallel documents. Based on this system and the data retrieved from Common Crawl, several websites were identified as possible candidates for crawling. He then implemented a database system that stores the crawled websites, the parallel documents found and other useful statistics like the number of documents crawled, and started a crawling process which has been applied to three different language pairs: English–Italian, English–German and English–French.

The web data crawling was finished as of February 2016 and it is currently being iteratively cleaned to ensure a greater quality and accuracy. The cleaning of the bilingual data is being done with a tool developed by ER1 called *TM Cleaner*[9]. In the case of the monolingual data collected, a set of *ad-hoc* scripts are being used. The monolingual data cleaning consists of eliminating the segments that contain tags or corrupt characters. Segments that do not have the correct language codes are accepted in the relevant monolingual corpora, once identified the appropriate language (e.g. the language code of the segment indicates that it is written in English, but it turns out to be German). A specific module has also been designed to detect and discard data containing inappropriate language.

As of August 2016, three cleaned monolingual corpora and three cleaned

[7]Here, we offer a summary of his approach and the latest developments in this work. For a more detailed explanation of the methodology, please consult *Deliverable D3.2. Multilingual corpus* [Barbu, 2016] of the EXPERT project.

[8]`http://commoncrawl.org/`

[9]For a description of this tool, see Section 3.1.2 in this Deliverable.

bilingual corpora have been collected. Their sizes are summarized in Tables 1 and 2. The monolingual data was obtained crawling websites in bilingual modality e.g. (English and German). As not all the crawled data was found to be truly parallel, not all downloaded documents could be mapped with translations. The German monolingual corpus, for example, consists of the union of all documents in German language detected in the websites crawled in the English–German modality.

Table 1: Size of the monolingual corpora collected by ER1.

Number of sentences		**Number of words**	
English: 26,483,977	Italian: 8,932,968	English: 699,761,128	Italian: 235,997,898
English: 81,008,261	France: 45,294,079	English: 2,490,577,768	French: 1,288,758,228
English: 57,464,786	German: 24,951,721	English: 1,669,062,885	German: 654,058,947

Table 2: Size of the bilingual corpora collected by ER1.

Language-Pair	**Number of Segments**	**Number of Words**
English-Italian	4,185,601	67,154,287/67,627,741
English -French	19,193,755	326,748,054/356,625,419
English-German	11,020,999	180,055,289/161,137,185

Both the monolingual and the bilingual corpora are currently being used in the MMT (Modern Machine Translation) European project[10]. The monolingual corpora are used for building language models and the bilingual corpora for building translation models.

2.3 Generation of artificial data

As discussed earlier, an alternative strategy for collecting new data is to artificially generate it. This strategy has been explored by ESR6 and ESR9 mainly.

ESR6's research focused on adding human feedback as the training data for a quality estimation (QE) system. To achieve that, she used the post-edited translations that human translators generated after analyzing the output of a MT system. Due to the scarcely available post-edited data and with the aim of training a high-quality QE system, Logacheva and Specia [2015a] proposed the generation of artificial data to train the QE model.

Logacheva and Specia [2015a] used the source segment and its human translation (target segment) to include different types of errors. For that,

[10] http://www.modernmt.eu/

the authors applied a set of operations at word level (insert, delete, replace and shift) to modify the original human translations. As a result of this process, they obtained new pairs of segments where the target sentence contained errors. The impact of the artificial data was tested on the sentence classification task, which aims at classifying the target sentences as good, almost good and bad. In their results, Logacheva and Specia [2015a] report that using this artificial data improved the sentence-level QE system performance and yielded a positive effect. However, when applied for word-level QE, the artificial data had a moderate impact on the system scores. The authors considered that these more moderate results may have been caused by a bad choice of the replacement words used to generate the artificial data.

ESR6 and ESR9 did a joint submission to the WMT15 word-level Quality Estimation shared task, in which they enhanced the training data with additional samples of artificial data [Logacheva et al., 2015]. The artificial data consisted of incomplete sequences of the already existing data, and the main goal was to enrich their QE model with new information. They used two different bootstrapping methods. The first method generated new sentences just using the first n words of each sentence and appending these new sentences to the training set. The second method concatenated to the training set all the trigrams of every sentence of the training corpus. In their results, they observed that the concatenation of incomplete examples to the training data proved to be particularly effective for training CFR models. More concretely, they report an improvement of the F1 score for the "BAD" class from 0.17 to 0.25 (i.e. $\sim 30\%$).

In Hokamp and Arora [2016], ESR9 and his co-author generated artificial data that was subsequently used in their SemEval 2016 submission to the Semantic Textual Similarity (STS) task. Inspired by the approach used by Ganitkevitch et al. [2013] to create the paraphrase database, they propose an innovative method for producing paraphrases. This approach combines domain-adaptation with paraphrase generation using two or more MT systems.

In their experiments for the SemEval STS task, they took the test set and translated it into Spanish and then back into English with an SMT system. The resulting "pseudoinstance" of the English original test sentence was then added to the data provided by the task organizers. Thus, the "paraphrases" added directly targeted the test data, which, as the authors explain, allows for "unsupervised domain adaptation of the model to the test data sets".

They validated whether this strategy of adding artificial data improves performance by running extra experiments using the 2015 Images dataset and comparing the performance of a system using their approach with the Paragram baseline, and with a model trained on task-internal data exclusively. The positive results obtained confirmed that generating artificial data can significantly improve performance.

3 Language technology, domain ontologies and terminologies

3.1 Translation Memories

Translation Memories (TMs) are a key element of any CAT tool and a crucial aid for translators and project managers alike. At the preparatory phase of any translation project, the TM is used to identify all segments previously translated as well as segments which are similar to the ones in the document to be translated and that can be used as a starting point of the translation to boost productivity. During the translation phase, translators save their translations in the TM and also use concordancers to search for fragments of segments in previous translations, terminology, etc. Within the EXPERT project, several innovative ways of improving TM systems have been explored.

3.1.1 Translation Memory Leveraging

Translation Memory leveraging is key for professional translators, as it determines the amount of segments that can be re-used in a new translation task. Given a segment to be translated, CAT tools look for such segment in the available TMs and provide the user with either an exact or approximate match from the TM, when available. As part of the EXPERT activities, ESR4 explored ways of improving the retrieval of fuzzy matches incorporating linguistic knowledge that resulted in several publications on this topic: Gupta et al. [2014], Gupta and Orăsan [2014] and Gupta et al. [2015][11].

Parra Escartín [2015] also explored ways of increasing the TM fuzzy match retrieval by fulfilling the translators' wishes. As a follow up of an internal survey carried out at Hermes, a shallow, language-independent method was implemented and tested in a pilot study where the requests of translators were integrated in the translation workflow.

Her strategy consists of generating new TM segments using formatting and punctuation marks. More concretely, all segments containing ellipses, colons, parentheses, square brackets, curly braces, quotations and text surrounded by tags were used as seed segments for creating new ones. The newly generated segments were modifications of the original ones, or fragments of them. In some cases, the strategy simply consisted of splitting in two an already existing segment.

The system was tested with a real translation project: a software manual written in English and to be translated into Spanish. The CAT tool used was

[11] For a more detailed report on ESR4's approach, see *Deliverable D6.4. Improved Hybrid Translation Tool* [Parra Escartín and Sepúlveda Torres, 2016b]. The user centered evaluation of the tool deployed by ESR4 is reported in *Deliverable D6.5. Final User Evaluation* Parra Escartín and Sepúlveda Torres [2016a].

memoQ 2015[12] and the impact of the approach was further tested using its "fragment assembly" functionality, a functionality that creates translations using fragments of the segments already existing in the TM.

The project had in total 425 segments accounting for 6280 words according to memoQ. Table 3 shows the project statistics using the project TM provided by the client. Additionally, memoQ identified 36 segments (418 words) which could be translated benefiting from its "fragment assembly" functionality.

TM match	Words	Segments
Repetitions	1064	80
100%	0	0
95–99%	4	2
85–94%	0	0
75–84%	285	14
50%–74%	2523	187
No Match	2404	142
Total	**6280**	**425**

Table 3: Project statistics according to memoQ using the project TM [Parra Escartín, 2015].

Three different TMs were used to further assess whether the size of the translation memory matters for generating translations of segments and retrieving more translations: the project TM, the product TM and the client TM. Table 4 summarizes the size of each of the TMs.

	Segments	Words		
		EN	ES	
Project TM	16,842	212,472	244,159	456,631
Product TM	20,923	274,542	317,797	592,339
Client TM	256,099	3,427,861	3,951,732	7,379,593

Table 4: Size of the different TMs used in Parra Escartín [2015].

The combination of the existing TMs with the newly generated segments decreased the number of segments not started in all cases (i.e. the fuzzy match retrieval improved), and increased the number of segments translated with the fragment assembly functionality. Moreover, the results showed that "the bigger the TM with new fragments, the higher the number of segments that benefit from fragments".

[12] https://www.memoq.com

3.1.2 Translation Memory Cleaning

One big innovation of the EXPERT project has been the organization of the first Automatic Translation Memory (TM) Cleaning Shared Task. This shared task was inspired by the work of ER1 on automatic TM cleaning presented in Barbu [2015] and the outcomes of the first workshop on Natural Language Processing for Translation Memories (NLP4TM), co-organized by two members of the EXPERT consortium: Dr. Constantin Orăsan (EXPERT Project Coordinator) and Dr. Rohit Gupta (ESR4) [Orăsan and Gupta, 2015].

A second edition of this workshop was organized by Dr. Constantin Orăsan, Dr. Carla Parra (ER2), Dr. Eduard Barbu (ER1) and Dr. Marcello Federico. The workshop was collocated with LREC 2016 [Orăsan et al., 2016], and as mentioned earlier, it included a shared-task on automatic TM cleaning. Translation companies rely on TM quality and curation to boost translators' productivity by maximizing TM reuse. However, more often than not, they face the need to deal with "dirty" TMs that need to be cleaned and from which wrong TM segments shall be deleted. The details about the preparation of the shared-task can be found in the relevant publication [Barbu et al., 2016]. As its name indicates, this shared task aimed at finding automatic ways of cleaning TMs that have not been properly curated and thus include incorrect translations.

For this first task bi-segments 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 TM segment represents a valid translation of its corresponding source. In particular, the following 3-point scale was 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 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 bi-segment 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 bi-segment was correct or incorrect. However, in contrast to the first task, a bi-segment was considered correct if it was labeled by annotators as ("1") or ("2"). Bi-segments 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 segments 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 total 6 teams participated in the shared-task[13]. ESR2, in cooperation with colleagues from the Jadavpur University, was a member of one of the participating teams (JUMT). In the shared task's website[14], there is an overview summary of the results[15], as well as the working notes submitted by all shared task participants [Ataman et al., 2016, Buck and Koehn, 2016, Mandorino, 2016, Nahata et al., 2016, Wolff, 2016, Zwahlen et al., 2016]

Besides the work on the shared task, ER1 released a tool to the community in September 2016[16]. The tool, called *TM cleaner*, is based on the code presented in Barbu [2015]. ER1 has re-written most of the code to make it easily usable by the translation industry and has additionally added new features:

1. Integration of the HunAlign aligner [Varga et al., 2005]. This component is meant to replace the automatic translation component as not every company can translate huge amounts of data. The score given by the aligner is smoothly integrated in the training model.

2. Addition of two modalities: the *train modality* and the *classify modality*. In the *train modality*, the features are computed and the corresponding model is stored. In the *classify modality* a new TM is classified based on the stored model.

3. Passing arguments through the command line. It is now possible, for example, to indicate the machine-learning algorithm that will be used for classification.

4. Implementation of hand written rules for keeping/deleting certain bilingual segments. These hand written rules are able to decide in certain

[13] A Journal paper summarizing the results of the shared task has been accepted for publication in the upcoming Special Issue of *Machine Translation*: *Natural Language Processing for Translation Memories*.

[14] http://rgcl.wlv.ac.uk/nlp4tm2016/shared-task/

[15] http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/results-1st-shared_task.pdf

[16] The tool is freely available on GitHub and includes a series of tutorials in its "Documentation" folder (https://github.com/SoimulPatriei/TMCleaner).

cases with almost 100% precision if a bilingual segment should be kept or not. This component can be activated/deactivated through an argument passed through command line.

5. Integration of an evaluation module. When a new test set is classified and a portion of it is manually annotated, the evaluation module computes the precision/recall and F-measure for each class.

The tool has been evaluated using three new data sets coming from aligned web sites and TMs. Moreover, the final version of the tool has been implemented on an iterative process based on annotating data and evaluating it using the evaluation module. This iterative process has been followed to boost the performance of the cleaner.

3.2 Taxonomy extraction

Tan et al. [2015a] explored ways of extracting hyperonyms and hyponyms by simplifying the means to extract a project matrix that transforms the hyponyms into their respective hyperonyms. Their strategy consisted of simplifying a previously complex process using a neural net to induce a hyperonym-hyponym ontology. Instead of the neural net, they used a word vector for the non-content word text pattern "is a". The authors presented their system to the SemEval Ontology Extraction shared-task and although it achieved modest results when compared against other participating teams, these were positive, which suggests that their approach could be used for other simple relation induction tasks between entities. To allow for replicability and reproducibility of the research reported in Tan et al. [2015a], the code was released as open source[17].

4 Learning from and informing translators

4.1 Quality Estimation

Quality Estimation (QE) is usually defined as the task consisting of automatically predicting the quality of some MT output without the existence of a reference [Blatz et al., 2004, Specia et al., 2009]. As such, QE is one of the most promising applications for the translation industry. Rejecting bad MT output at an early stage is crucial to avoid delaying professional translators in post-editing tasks and to ensure that the MT output given to post-editing complies with a certain quality threshold that will boost their productivity.

QE systems combine language-independent or language-dependent features of source and target texts. These texts usually include a few translation

[17] https://github.com/alvations/Endocentricity

samples tagged for quality by humans and they are used to build a machine learning algorithm to predict the quality of unseen translation data [Blatz et al., 2004, Specia et al., 2009]. Feature engineering plays a key role in QE and leads to a potential bottleneck, because including a set of features is an expensive process and the impact of the features on MT performance vary across datasets and language pairs [Scarton and Specia, 2014b, Shah et al., 2015].

Several researches have focused on quality estimation at word-level, phrase-level, sentence-level, and document-level. Within the EXPERT project, a variety of approaches of QE at different levels have been addressed too.

As part of the EXPERT activities, ESR6 explored ways of collecting and extracting useful human feedback to improve statistical machine translation systems.

Logacheva and Specia [2014a] used active learning (AL) techniques for extracting the training data to improve a machine translation baseline system. The novelty of this proposal was to add real post-editions for the MT training data. The active learning strategies were based on quality predictions to select the sentences which could improve the performance of an MT system. The authors assumed that "the most useful sentences are those that lead to larger translation errors" [Logacheva and Specia, 2014a]. For QE training, they used the HTER scores predicted by a QuEst [Specia et al., 2013] system trained with MT output and its corresponding post-edited version.

Their approach works as follows: In the first step, a baseline MT system was trained for an initial training data. After that, the AL strategy is activated to select a batch of 1,000 sentences for a different data set. The selected sentences are transferred to the MT training data and the MT system is rebuilt with the original data and the new batch of sentences. This process is repeated until the AL data is empty. Logacheva and Specia [2014a] implemented the following AL strategies:

- QuEst: Logacheva and Specia [2014a] AL proposed;
- Random: random selection of sentences;
- HTER: oracle-based selection based on true HTER scores of sentences, instead of the QuEst estimated HTER scores.; and
- Ananthakrishnan et al. [2010] AL strategy.

The Logacheva and Specia [2014a] AL strategy outperforms the Ananthakrishnan et al. [2010] and Random strategies, but the HTER method performs better than the Logacheva and Specia [2014a] approach. However, the authors perceived a tendency of QuEst to assess long sentences as having lower quality. As a consequence, Logacheva and Specia [2014a] reported that this bias significantly better influences the performance of the MT system than the HTER method, because QuEst adds more data to train the

MT system. Logacheva and Specia [2014a]'s experiments show that adding post-editing results significantly improves the BLEU score of MT.

Logacheva and Specia [2014b] extended the experiments proposed by Logacheva and Specia [2014a] for extracting training data to improve the MT system. In this work, post-editions and human translations of source sentences were included to train an MT system. The authors also used the human post-editions to post-edit the sentences that were selected through the QuEst strategy for having low quality.

Logacheva and Specia [2014b] conducted two sets of experiments to analyze the advantages of using post-edition of MT instead of human translations. For the first experiment, human translators manually translated the source sentences. In the second one, an MT system was used to translate the source sentences, and then a human translator post-edited the set of translations. These post-editions composed the target side of the new parallel dataset.

The results in Logacheva and Specia [2014b] showed that the improvement obtained with the system trained with post-editing data was substantially higher than that obtained by a system with additional human translations. Following these results, the authors suggested that it is possible to "reduce the translator's effort in creating data for active learning while getting even better improvements in the quality of the resulting MT system" [Logacheva and Specia, 2014b].

Considering the intuition that the statistical MT engines are phrase-based, Logacheva and Specia [2015b] explored an approach to assess QE of MT at phrase-level. The authors assumed that predicting the quality of a sequence of words could provide additional information, which is not available at word-level. This additional information could also be used to guide the MT decoding process to avoid some errors [Logacheva and Specia, 2015b].

In order to build a phrase-level QE system, Logacheva and Specia [2015b] proposed two segmentation methods based on the segmentation produced by an SMT system. The first method used the MT system which generated the translation to reproduce the original source and target segmentation. The second method segmented the source sentence, combining only the target segmentation and the alignments between source and target.

Logacheva and Specia [2015b] used datasets with post-edited MT that have been labelled at the word-level (as "BAD" or "OK"), such as the training model tagging the test phrases. They combined the labels of each word into the phrase following three categories:

- optimistic: if half or more words in the phrase have the label "OK";
- pessimistic: if 30% or more words in the phrase have the label "BAD"; and
- super-pessimistic: if any word in the phrase has the label "BAD".

Logacheva and Specia [2015b] combined 79 features produced by the QuEst QE framework and other features based on representations of words. Theses last features were obtained only for the source sentences and with a word2vec[18] tool. The authors explored the CRF model and a Random Forest classifier to build the phrase-based prediction model.

Logacheva and Specia [2015b] compared their approach with other word-level systems which were presented on the WMT-14 and the WMT-15 QE shared tasks. The authors show how their system outperforms the other ones and, more particularly, they show that:

- sentence-level features achieved better results than word-based features;
- the CRF model was better that the Random forest; and
- the best tagging strategy was classifying as "BAD" the phrases in which at least one word had been labeled as "BAD".

At the end, Logacheva and Specia [2015b] confirmed that "sentences with less errors do not contribute much for word-level QE."

ESR6 also researched alternatives to include unsupervised features to improve the QE task at different levels. Shah et al. [2015] investigated the use of word embeddings, or word vector representations and neural networks, in both sentence-level and word-level QE. The authors proposed strategies to extract features from such resources and use them to learn prediction models. Shah et al. [2015] combined the features extracted for the above resources with a set of other features obtained through the QuEst QE framework.

Shah et al. [2015] used the neural network Continuous Space Language Model (CSLM) to obtain features for sentence-level QE. The input of the language model was the context of the prediction $h_j = w_{j-n+1}, ..., w_{j-2}, w_{j-1}$ to estimate the posterior probabilities of all words of the vocabulary: $P(w_j|h_j)$ [Shah et al., 2015]. The CSLM models were trained with the CSLM toolkit[19]. The authors extracted the probabilities for the source and its translation and applied them as independent features. They also combined them with other features.

Additionally, Shah et al. [2015] trained the word-embedding model proposed by Mikolov et al. [2013] using large monolingual data "with a neural network architecture that aims at predicting the neighbours of a given word" [Shah et al., 2015]. In contrast to the above language model, in this case the neural network architecture predicts the word in the middle of a sequence, given the representation of the surrounding words [Shah et al., 2015]. Inspired by the Mikolov et al. [2013] approach, the authors also added the source-to-target similarity score feature that was obtained mapping the source and target vector through a dictionary.

[18] https://code.google.com/p/word2vec/
[19] http://www-lium.univ-lemans.fr/cslm/

In general, Shah et al. [2015] considered that their results are quite promising, especially those produced by the word-embeddings model, "because, although it uses only features learned in an unsupervised fashion, it was able to outperform the baseline as well as many other systems."

On the other hand, ESR7 concentrated efforts on QE at document-level. ESR7's research has focused on finding the best features for document-level QE and addressed the phenomena of discourse in QE. ESR7's research hypothesis is that "discourse information can be used to improve state-of-the-art QE models by detecting issues related to discourse due to the way machine translations are produced".

In this scenario, Scarton and Specia [2014a] explored discourse features and pseudo-reference translations to estimate the translation quality at document-level. The discourse features included by the authors were Lexical Cohesion (LC) and Latent Semantic Analysis (LSA) Cohesion. The LC features were based on counting the repetitions of tokens, lemmas or nouns included in the document. On the flip side, to compute LSA Cohesion features, Scarton and Specia [2014a] created a matrix of word frequency per sentence, applied Singular Vector Decomposition to the matrix and calculated the correlations between the word vectors of each sentences.

Scarton and Specia [2014a] also included features based on pseudo-references. "Pseudo-references are translations produced by one or more external MT systems, which are different from the one producing the translations" [Scarton and Specia, 2014a]. In this experiment, the authors estimated BLEU and TER scores between the output of the interest MT system and the pseudo-reference outputs, produced by MOSES, SYSTRAN and GOOGLE. These set of measures were used as features to train the quality prediction models [Scarton and Specia, 2014a].

Scarton and Specia [2014a] tested the impact of the proposed features on the quality of translation and results showed significant improvements when pseudo-reference features were included. However, LC and LSA Cohesion were decisive in improving the baseline performance [Scarton and Specia, 2014a].

Additionally, Scarton and Specia [2015] guided an extensive study regarding the correlation between several discourse features and HTER for document-level quality estimation. Two corpora of English and French were used in the quantitative analysis, which included the following discourse phenomena:

- Lexical Cohesion;
- LSA Cohesion;
- Counts of connectives;
- Counts of pronouns;
- Discourse unit segmentation;
- Rhetorical Structure Theory (RST); and

- Elaboration relation.

Because some of the tools used by Scarton and Specia [2015] to extract discourse features were only available for English, the evaluations were conducted for English source and target texts. Results show a significant correlation between the proposed discourse features and the HTER score. These correlations were higher than the correlation of the features based on sentence-level quality estimation.

ESR7 and coauthors participated in the WMT15 shared task on document-level quality estimation, specifically on the paragraph-level QE task. In their submission, Scarton et al. [2015] proposed strategies to select the best features to deal with the QE task. The authors combined a set of features based on sentence-level QE that appeared on previous works and features based on discourse information and presented by Scarton and Specia [2014a].

To obtain the best set of features, Scarton and Specia [2014a] explored two feature selection algorithms: backward based on Random Forests and exhaustive search. The first method consists of eliminating features with a lower position in the rank, until the set of features is obtained which better fits the predictions. On the other side, the second algorithm was only applied to select the features included on the baseline features set. Scarton and Specia [2014a] explored the "efficacy of the baseline features by learning one Bayesian Ridge classifier for each feature and evaluating the classifiers based on Mean Absolute Error". The authors also implemented an exhaustive feature selection search to rank all possible feature combinations among the baseline features.

The authors explored their approach on the language pair of English and German and presented results for both language directions. An important finding of this experiment was that using discourse features does not mean that there were significant changes on the baseline performance. However, the authors linked this result to the fact that the shared task data was composed by short paragraphs, and in texts with these characteristics the discourse features are becoming less effective.

The varieties of features and algorithms explored by ESR7 and coauthors regarding QE were included on QUEST++ [Specia et al., 2015], an expanded version of the open source sentence-level toolkit, QUEST [Specia et al., 2013]. QUEST++ can predict the quality for texts at word, sentence and document level. The tool provides a pipelined processing, that allows the interaction between the three levels: word, sentence and document QE.

QUEST++ comprises several features and machine learning algorithms to build and evaluate models for quality estimation. The tool is composed by two principal modules: a feature extraction module and a machine learning module. The feature extraction module requires pairs (source and target) of raw text files and some resources, like MT corpus and language models for source and target language. QUEST++ used a set of 40 features of seven

general types [Specia et al., 2015].

- Target context: Explore the context of the target word.
- Alignment context: Explore the word alignment between source and target sentences.
- Lexical: Explore POS information on the source and target words.
- Language Model: Consider the n-gram frequencies of a word's context with respect to an LM.
- Syntactic: Use the Stanford Parser for dependency parsing.
- Semantic: Explore the polysemy of source and target words.
- Pseudo-reference: Explore the similarity between the target sentence and a translation for the source sentence produced by another MT system.

On the other hand, QUEST++ machine learning module provides different regression and classification algorithms, feature selection methods and grid search for hyper-parameter optimisation to deal with the QE task [Specia et al., 2015]. The machine learning module contains some scripts to interface the Python toolkit scikit-learn[20] [Pedregosa et al., 2011] and interacts with the feature extraction module to create the QE model.

Finally, Scarton et al. [2016] explored a set of features, used on their previous works with machine learning techniques to build the following two systems, which were submitted to the WMT16 QE shared task.

- GRAPH-BASE: Combined discourse features, like counts on pronouns, connectives and RST relations with the official baseline features to train a Support Vector Regression (SVR) algorithm.
- EMB-BASE-GP: Combined word embeddings from the source documents with the official baseline features to train a Gaussian Process with two-kernels: one for word embeddings and one for baseline features.

Scarton et al. [2016] trained the same word-embeddings model presented by Shah et al. [2015]. Some details of this model were explained earlier. To train the model, the authors used Google's billion-word corpus for English and a combination of WMT Europarl, News-commentary and News-crawled for Spanish. Scarton et al. [2016] built the document word-embeddings by averaging word-embeddings in the document and used these representations as features.

The model that combined word embeddings with baseline features (EMB-BASE-GP) outperformed the model that included discourse information (GRAPH-BASE). With a 0.391 Pearson r correlation score, EMB-BASE-GP was the winning model of the scoring sub-task. However, the authors

[20] http://scikit-learn.org/stable/

assumed that the GRAPH-BASE low performance was related with the fact that the discourse tools used to extract the feature were only available for source texts (English) and not for the target text (Spanish).

Within the EXPERT project other ESRs, such as ESR9 and ESR12 also contributed to QE on the MT context.

Béchara et al. [2016] explored ways of integrating Semantic Textual Similarity (STS) features into Quality Estimation using the MiniExperts STS tool developed by Béchara et al. [2015]. The authors use semantically similar sentences to the ones that have been translated by an MT system together with their similarity scores as features to estimate the quality of the new translations.

Their approach works as follows: For each sentence A, for which MTQE is needed, they retrieve a semantically similar sentence B which has been machine translated and has a reference translation or a quality assessment value. Then, they extract three scores:

- The STS score, which represents the STS between the source sentence pairs;
- The quality score for sentence B, which is either S-BLEU score based on a reference translation, or a manual score provided by a human evaluator; and
- The S-BLEU Score for Sentence A, which is computed using Sentence B as a reference.

The authors test their approach with different data sets and obtain encouraging results. This suggests that the three STS features they added to the MTQE system can yield better results when a sufficiently similar sentence against which to compare is available.

In addition, ESR9 participated on the SemEval-2016-Task 1: Semantic Textual Similarity (STS) and presented a model to predict semantic similarity scores between unseen source and target pairs of sentences [Hokamp and Arora, 2016]. The authors used a set of sentences annotated with their semantic similarity scores (0, no relation to 5, semantic equivalence), as a training set.

Hokamp and Arora [2016] experimented models which learn sentence-level embeddings considering simple bag-of-words averages of the embeddings in each sequence. The authors combined the Paraphrase Database (PPDB) [Wieting et al., 2015] to train embeddings, and tuned these embeddings with the official training data available for STS task. Hokamp and Arora [2016] also generated artificial paraphrases using machine translation and a similarity score obtained through embedding networks as the predicted score.

The authors perceived good performance on the STS task using only the proposed embedding models. However, Hokamp and Arora [2016] observed

an improvement in performance after combining traditional features with the similarity score learned by the proposed embedding models.

5 Hybrid corpus-based approaches

Prior to EXPERT, "hybrid corpus-based solutions considered each approach individually as a tool, not fully exploiting integration possibilities". The proposed EXPERT solution was to "fully integrate corpus-based approaches to improve translation quality and minimize translation effort and cost."

Several ESRs have worked on this topic within the EXPERT project. In what follows, we present a summary of their work, subdivided into two major categories: Hybrid Statistical Machine Translation (Section 5.1), and Automatic Post-Editing (Section 5.2).

5.1 Hybrid Statistical Machine Translation

In recent years, a lot of research has been done with the aim of integrating linguistic information into Statistical Machine Translation (SMT) systems. Within EXPERT, several researchers have also explored ways of integrating linguistic knowledge into STM. While some researchers have focused on ways of reducing the number of Out-of-Vocabulary words, others focused on improving lexical selection, improving word alignment, integrating Translation Memories into SMT, domain adaptation and yielding better results for morphologically rich languages. In the following subsections, we report on this work.

5.1.1 Improving vocabulary coverage and accuracy

When an SMT system faces an unseen word, it fails to produce a translation of it. This phenomenon is referred to as the problem of Out of Vocabulary words (OOVs) and it constitutes a challenge for SMT systems. One strategy of dealing with OOVs relies on the usage of terminologies to increase the number of seen words at the training stage and thus automatically reduce the number of OOVs. This strategy is also used to improve word alignment. By concatenating glossaries to the training corpora of SMT systems, the probabilities of the terms included in the glossaries ending up aligned with their translations potentially increase.

Tan and Pal [2014] explored how to integrate terminology in their WMT 2014 submission. Their system, called *manawi* (Multi-word expression And Named-entity And Wikipedia titles), added automatically extracted bilingual MWEs and NEs as additional parallel data to training data. Moreover, the data was cleaned in a pre-processing step based on sentence-alignment features. For them, a sentence pair is parallel if the ratio between the number of characters in the source and the target sentence are coherent with

the global ratio of the number of source-target characters in a fully parallel corpus. To compute this, they first measured the global mean ratio of the number of characters of the source to the target sentence and they then filtered out all pairs exceeding or below 20% of the global ratio. Their strategy yielded promising results indicating that pre-processing and inclusion of MWEs and NEs may improve the results obtained in SMT systems.

In Tan et al. [2015b], the authors explored different ways of adding dictionaries to SMT systems. More concretely, they explored the impact of using dictionaries in the two traditional ways within SMT:

(a) Appending the dictionary to the training data and treat it as part of the parallel data. This assumes that by doing so the words contained in the dictionary will increase their probability of ending up aligned together and thus the system will learn better translations. The authors call this approach the "passive" approach.
(b) Forcing the translation of the dictionary entries during decoding, which the authors name the "pervasive" approach.

The authors additionally test how many times a dictionary shall be appended to maximize the overall MT evaluation scores in terms of BLEU. They also test the same when the "passive" approach is used in combination with the "pervasive" one. Table 5 summarizes their results. For their experiments, they focused on the English→Japanese language pair. As may be observed, the best results are obtained by appending the dictionary twice to the training data ("passive" approach), or adopting the combination of the "passive" and the "pervasive" approach, in which case it is necessary to append the dictionary four times to the training data to obtain statistically significantly better results than the "passive" approach in isolation.

	- Pervasive	**+ Pervasive**
Baseline	16.75	16.87
Passive x1	16.83	17.30
Passive x2	**17.31**	16.87
Passive x3	17.26	17.06
Passive x4	17.14	**17.38**
Passive x5	16.82	17.29

Table 5: BLEU scores obtained by Tan et al. [2015b] using the passive and pervasive approaches tested.

Pal et al. [2014c] report on the participation of ESR2 and other colleagues in the NLP Tools Contest of the International Conference on Natural Language Processing (ICON 2014). In their submission, Pal et al. [2014c] explore different strategies to beat the baseline system. Their main innovations are the following:

1. Effective pre-processing.
2. Use of explicitly aligned bilingual terminology (in their case named entities).
3. Simple but effective hybridization technique for using multiple knowledge sources.

They report that their "hybrid system potentially improves over the baseline statistical machine translation performance by incorporating additional knowledge sources such as the extracted bilingual named entities, translation memories, and phrase pairs induced from example-based methods." Their best system consisted of merging all approaches into a combined one.

This approach is further explored in Pal et al. [2015a], where the authors also propose a hybrid system which incorporates additional linguistic knowledge in the form of extracted bilingual named entities and chunk pairs. The additional knowledge is extracted in a pre-processing step which additionally cleans and clusters sentences based on sentence length.

In the case of NEs, they first identify the NEs in both languages. They use the Berkeley NE Recognizer for English and an implementation of the NE described in Ekbal and Bandyopadhyay [2009] for Bengali. Once identified, they join the NEs by means of an underscore ("_") to emulate single tokens. Then they isolate the sentences containing NEs and subsequently align the NEs. If the sentence is just an NE, the alignment is straightforward. Else, they use the transliteration method proposed by Ekbal et al. [2006] as a strategy for aligning them.

In the case of chunk alignment, chunks are extracted from the source side of the corpus using the Stanford Part-of-Speech tagger [Toutanova et al., 2003][21]. In the case of Bengali, a shallow parser is used. The head of each chunk is also identified. If the heads of the chunks are translations of each other, the chunk is aligned.

Despite all the authors' efforts, a simple confusion network-based system combination outperforms all the individual MT systems that the authors tried. They argue that one possible problem may have been that tuning was based on BLEU, but further research is needed to verify this hypothesis.

Pal and Kumar Naskar [2016] build up the exploration of chunk word alignments and explore hybrid methods for word alignment. The authors propose a system which combines the outputs of three different word aligners: Giza++ [Och and Ney, 2003]; the Berkeley aligner [Liang et al., 2006]; and a rule-based aligner of their own which aligns Named Entities (NEs) and chunks for the language pair English→Bengali. They run several SMT experiments using their hybrid word alignment approach and obtain substantial improvements (10.25 BLEU points absolute, 93.86% relative) over the baseline.

[21] http://nlp.stanford.edu/software/tagger.shtml

Finally, in Pal et al. [2016a] the authors merge the hybrid word alignment approach with Forest to String Based Statistical Machine Translation (FSBSMT). FSBSMT is a forest-based tree sequence to string translation model which is used in some syntax based statistical machine translation systems. The model automatically learns tree sequence-to-string translation rules from a given word alignment estimated on a source-side-parsed bilingual parallel corpus. The authors report very positive results when combining these two strategies. Their results seem very promising, as they achieve considerable improvement over state-of-the-art Hierarchical Phrase-Based SMT. The integration of NEs and their translations yield further improvements in the system output. In their experiments, involving the English→Bengali language pair, they achieve a 78.5% relative (+9.84 BLEU points absolute) improvement over the hierarchical phrase-based SMT baseline.

5.1.2 Integrating Translation Memories in SMT

Li et al. [2014, 2016a] and Li et al. [2016b] explore ways of integrating Translation Memories (TMs) into Statistical Machine Translation (SMT). In Li et al. [2014], the authors present a discriminative framework which allows for the integration of TMs into SMT. They add to a phrase-based SMT system several TM feature functions that model the relationship between the source sentence and the TM instances. After incorporating the TM features to their system, they retrieve significantly better results than a phrase-based SMT system used as a baseline both in English→Chinese and English→French translation tasks. They additionally propose a way to efficiently use multiple fuzzy matches, and their experiments yield significant results that are even better. Moreover, although the approach proposed by Li et al. [2014] uses most of the features proposed by Wang et al. [2013], their methodology is simpler and yet obtains comparable results.

In Li et al. [2016a], the authors explore further the integration of TMs and SMT. In this work, the authors combine a TM with a syntax-based MT system. The combination is done using sparse features extracted during decoding and such features are based on translation rules and their corresponding patterns in the TM. Their approach is tested on real industrial data with very promising results, particularly because the improvements were achieved over already high baseline systems (62.8 BLEU, 79.5 METEOR and 26.5 TER in the case of a state-of-the-art phrase-based MT system). They report statistically significant improvements of up to +3.1 BLEU, +1.6 METEOR and -2.6 TER. The authors also explore the fuzzy matches that yield better overall scores and find out that these seem to be obtained for the highest fuzzy match bands.

Finally, in Li et al. [2016b], the authors explore how to combine TMs and SMT at a sub-sentential level. Their approach consists of using a "con-

strained word lattice, which encodes input phrases and TM constraints together, to combine SMT and the TM at phrase level". The TM constraints come from two sentence-level constraint approaches: addition [He et al., 2010] and subtraction [Koehn and Senellart, 2010]. The authors run English→Chinese and English→French SMT experiments and obtain significantly better results than previous combination methods such as sentence-level constrained translation and phrase-level combinations.

5.1.3 Dealing with syntax and morphology in SMT

Translation into morphologically rich languages often proves to be problematic under the assumption of current models. There are three major challenges for SMT systems in this area:

1. Freer word order is difficult to model.
2. The usually large space of possible morphological inflections leads to problems of data sparsity.
3. Morphological agreement is often expressed over long distances.

For many language pairs, both syntax and morphology have to be taken into account for producing a good translation and both may interact in various ways. The focus of ESR10's research within EXPERT has been to investigate and produce statistical models for MT that compose translation units in a better way.

In order to better manage morphology, word order and their interaction, Daiber and Sima'an [2015b] proposed a model for projecting the target morphological features on the source string and its predicate-argument structure. In order to fit the morphological feature set to training data, they proposed the usage of a latent variable model that learns the feature set. The assessment of the learned feature set showed that it achieved a quality comparable to that of a manually selected set for German. The morphology is then integrated into the back-end phrase-based model so that it can also be trained on projected features and thus provide a more efficient pipeline.

In their experiments, they show that it is possible to bridge the gap "between a model trained on a predicted and another model trained on a projected morphologically enriched parallel corpus". In fact, the results obtained demonstrated that the projection of a small subset of morphological attributes to the source side can yield major improvements. It also has the additional advantage of reducing the complexity of predicting such features.

To deal with the rich set of word order choices that is typical of morphologically rich languages, Daiber and Sima'an [2015a] explored a model based on the popular preordering algorithm proposed by Lerner and Petrov [2013]. In their work, Daiber and Sima'an [2015a] proposed a novel preordering model that "is able to produce both n-best word order predictions

as well as distributions over possible word order choices in the form of a lattice". Following the authors, this has the advantage of making the model a good fit for richer models that tackle both syntax and morphology.

Daiber and Sima'an [2015a] evaluated their methodology by running SMT experiments using a common system setup and obtained very positive results. They also observed that the preordering quality of the English–German language pair is improved by integrating non-local language model features via cube pruning.

Additionally, Daiber et al. [2016] studied the influence of word order freedom and preordering in SMT. They did an empirical comparison of distant language pairs (English*to*German and English*to*Japanese) with the aim of determining the difficulty of predicting the word order of the target language's word order based on the source's one. Their results confirmed that addressing uncertainty in word order predictions by means of permutation lattices "can be an indispensable tool for dealing with word order in machine translation". They further developed this method by introducing a lattice-preordered input for training SMT systems and found out that lattices are crucial for languages with a freer word order like German, while they are also helpful for other language pairs for which reordering already has proven to be useful.

Pal et al. [2014a] also explored ways of improving SMT results by re-ordering the source sentences in such a way that they mimic the target language word ordering prior to training. This re-ordering was based on chunk word alignments and its main advantage relies in that it does not require syntax parsers or complex language processing tools other than a chunker for the source language. In their experiments, they obtain statistically significant improvements when compared with tree-based or simple word alignment re-ordering methods. Moreover, manual evaluation of the output revealed that this method also retrieves statistically significant improvements in terms of word alignment.

Finally, besides dealing with morphology and syntax, a problem to be dealt with in the case of language pairs involving Germanic languages such as German, is compounding. In these languages, compounding is productive, and thus effective methods to automatically identify compounds and deal with them in SMT are required. As part of his work in the EXPERT project, ESR10 introduced an unsupervised method based on regularities in the semantic representations of words.

This method is presented in Daiber et al. [2015] and exploits the regularities in the semantic word embedding space to model the composition of compounds based on analogy. After identifying that the semantic vector space lends to modeling compounds based on their semantic head, they extracted compound transformations using the approach proposed by Sori-

cut and Och [2015][22]. Their proposed algorithm applies these structures to compound splitting and the results obtained in their experiments prove that their system outperforms the traditionally used compound splitters based on shallow frequencies. Moreover, their system proved to be particularly good for splitting highly ambiguous compounds, and when used as part of the preprocessing pipeline of SMT systems, their system also showed positive results when compared with the traditional compound splitting method. The implemented tool is freely available under an Apache license on github[23].

5.1.4 Domain Adaptation for SMT

Domain adaptation methods aim at tuning for a particular domain an already existing SMT engine that was trained using out-of-domain data. Usually, this is achieved by adding a small in-domain data sample which is available for carrying out the domain adaptation [Hoang and Sima'an, 2015]. ESR11's research has focused on this particular topic achieving significant progress. One of the major advances obtained has been in improving the word alignment statistics with respect to several domains. The work by ESR11 has been guided by the following research question: "Given domain information for the same small subsets of a large mixed corpus, how can we use the information as priors and learn the corresponding domain-conditioned word alignment statistics for the pool of sentence pairs in the heterogeneous data?".

Hoang and Sima'an [2015] proposed a latent domain HMM word alignment model to produce a better alignment in corpora with several domains. The model induced domain-conditioned probabilities for each sentence pair into the corpus and was trained considering a small seed of samples from different domains. Hoang and Sima'an [2015] reported a large number of experiments over three different size corpora of English and Spanish. The experiments showed that:

- The latent domain model produced significant improvements in word alignment accuracy over the original HMM alignment model;
- The latent domain model was more sensitive to the domain than the original HMM alignment model; and
- After integrating the latent domain model on different translation tasks, an improvement of the resulting SMT system was observed.

Additionally, Hoang and Sima'an [2014] addressed the problem of creating adequate training data (set of sentence pairs) to train a machine translation model. The authors carried out experiments in data selection and

[22] An implementation of this approach was done by ESR10 and it is freely available at https://github.com/jodaiber/vec2morph.

[23] https://github.com/jodaiber/semantic-compound-splitting

proposed a model to estimate the probability of a pair of sentences to belong or not to belong to a specific domain. To build a model, Hoang and Sima'an [2014] trained an EM algorithm based on the in-domain corpus statistics as a prior.

The authors tested their approach on English→Spanish corpora and proposed different ways to model the importance of the sentence pairs in a corpus with a varied set of domains. The data selection model proposed by Hoang and Sima'an [2014] yields better performance than the baselines used for comparison purposes. The authors trained an SMT system using their proposed data selection model and appreciated an improvement in the system performance.

Cuong et al. [2016] also presented an unsupervised method to adapt an MT system to all domains. The proposed method does not require a knowledge of the target domain, and thus it may be applied in a different scenario of domain adaptation [Cuong et al., 2016]. The model was trained to induce subdomains and integrate a set of features to estimate the importance of a specialized phrase in an induced subdomain. Cuong et al. [2016] tested their approach on three language pairs (English→French, English→German, and English→Spanish) and explored adaptation to 9 domain adaptation tasks. They report a modest, but consistent improvement in BLEU, METEOR and TER and they show that their method outperforms a set of proposed baselines.

5.2 Automatic Post-Editing

Automatic Post-Editing (APE) is the task of automatically attempting to correct Machine Translation (MT) output so that it is more similar to a human reference. Sometimes, these systems are used as a post-processing module in MT tasks to improve the overall performance of an MT system. One approach towards APE consists of training an MT system with raw MT output (used as the source language), and its corresponding human post-edits (used as the target language). Such an APE system thus "translates" an MT output into its corresponding corrections. Within the EXPERT project, ESR2 has explored different ways of improving APE systems.

In Pal et al. [2015c], ESR2 and his co-authors explored the impact of hybrid word alignment in APE tasks. To do so, they integrated in their APE system an edit-distance based monolingual aligner. Their system ranked third out of the seven submissions to the 2015 APE shared task and slightly outperformed the baseline system. Their work demonstrates that hybrid word alignment can play a crucial role in APE tasks.

Pal et al. [2016b] present an APE system based on a bidirectional recurrent neural network (RNN) model. It consists of an encoder that encodes an MT output into a fixed-length vector from which a decoder provides a post-edited (PE) translation. The results obtained showed statistically significant

improvements over the original MT output (+3.96 BLEU), a phrase-based APE system (+2.68 BLEU) and a hierarchical one (+1.35 BLEU). The original MT system was Google Translate (English→Italian), and the baseline system was already obtaining an impressive 61.26 BLEU score, which makes it more difficult to beat.

Besides the automatic evaluation, Pal et al. [2016b] run a human evaluation of the output. This revealed that their system drastically reduces the preposition insertion and deletion error in the Italian Google Translate output, and additionally handles the improper use of prepositions and determiners and reduces word ordering errors to some extent.

Finally, in Pal et al. [2016c], ESR2 and their co-authors tested with positive results an Operation Sequential Model (OSM) combined with a phrased-based statistical MT (PB-SMT) system. Their APE system was trained on the MT outputs (TL_{MT}) produced by a black-box MT system and their corresponding post-edited version (TL_{PE}). When evaluated against the official test set of the APE shared task, their system achieved a +3.2% relative improvement in BLEU over the raw MT output (+1.99 absolute points). In the case of TER, they achieved -0.66 absolute points and -0.25% relative improvement.

6 Conclusion

In this Deliverable we have reported on the work related with corpus-based Approaches carried out within the EXPERT project. As may be observed, the contributions of the project to state-of-the-art approaches are manifold.

Researchers have worked extensively in data collection efforts exploring new ways of compiling parallel and comparable corpora, and new tools have been released to the community or will be released by the end of the project. As far as Language Technologies are concerned, efforts have been made to develop tools to improve Translation Memory Leveraging and reuse of already existing Translation Memories. Moreover, ways of automatically curating and cleaning Translation Memories have also been explored and a new tool will be released to the community in September 2016.

Bearing in mind the current trend of using Machine Translation to boost translators productivity and alleviate the time and costs related to translation tasks, Quality Estimation (i.e. determining the quality of MT output without a given reference) will play a crucial role in the short time. Researchers in EXPERT have also contributed to this field and worked to bring the state-of-the-art to a new level that will allow translators to use QE in their MTPE workflow to discard bad MT output as soon as possible.

Besides attempting to predict the quality of MT output, several EXPERT researchers have focused on improving the quality of MT systems themselves. We have summarized how many of them have focused on the

hibridization of SMT systems to bring linguistics into SMT and yield better output quality. Besides work on broadly researched languages such us English–Spanish, several researchers in EXPERT have contributed toward the improvement of SMT in minority languages, rarer language pairs (e.g. English–Bengali) and morphologically rich languages such as German.

Last but not least, research on Automatic Post-Editing has also shown promising results and could easily be integrated in the translation industry, thus helping translators to post-edit faster because recurring errors will have been corrected by the APE system beforehand.

EXPERT aimed at bringing the state-of-the-art in Machine Translation and Translation Technologies to a new level. The work summarized here, as well as in the two other final Deliverables, *D6.4 Improved Hybrid Translation Tool* [Parra Escartín and Sepúlveda Torres, 2016b] and *D6.5 Final User Evaluation* [Parra Escartín and Sepúlveda Torres, 2016a], is a great proof thereof.

References

Steven Abney and Steven Bird. Towards a data model for the universal corpus. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 120–127, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-015. URL http://dl.acm.org/citation.cfm?id=2024236.2024257.

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. Discriminative Sample Selection for Statistical Machine Translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 626–635, Cambridge, MA, October 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D/D10/D10-1061.pdf.

Duygu Ataman, Masoud Jalili Sabet, Marco Turchi, and Matteo Negri. 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.

Eduard Barbu. Spotting false translation segments in translation memories. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 9–16, Hissar, Bulgaria, September 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-5202.

Eduard Barbu. Deliverable D3.2: Multilingual corpus. Deliverable in the EXPERT project, 2016. URL http://expert-itn.eu/sites/default/files/outputs/d3.2_multilingual_corpusexpert_d3.2_20160128.pdf.

Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Marcello Federico, Luca Mastrostefano, and Constantin Orăsan. 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 2016.

Hanna Béchara, Hernani Costa, Shiva Taslimipoor, Rohit Gupta, Constantin Orăsan, Gloria Corpas Pastor, and Ruslan Mitkov. MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In *9^{th} Int. Workshop on Semantic Evaluation*, SemEval'15, pages 96–101, Denver, Colorado, June 2015. ACL.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (CoLing-2004)*, pages 315–321, 2004.

Christian Buck and Philipp Koehn. 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.

Hanna Béchara, Carla Parra Escartín, Constantin Orăsan, and Lucia Specia. Semantic Textual Similarity in Quality Estimation. *Baltic Journal of Modern Computing*, 4(2):256–268, June 2016.

Hernani Costa. Deliverable D3.1: Famework for Data Collection. Deliverable in the EXPERT project, 2015a. URL http://expert-itn.eu/sites/default/files/outputs/expert_d3.1_20150213.pdf.

Hernani Costa. Assessing Comparable Corpora through Distributional Similarity Measures. In *EXPERT Scientific and Technological Workshop*, pages 23–32, Malaga, Spain, June 2015b.

Hernani Costa, Gloria Corpas Pastor, and Ruslan Mitkov. Measuring the Relatedness between Documents in Comparable Corpora. In *11^{th} Int. Conf. on Terminology and Artificial Intelligence*, TIA'15, pages 29–37, Granada, Spain, November 2015a.

Hernani Costa, Gloria Corpas Pastor, Miriam Seghiri, and Ruslan Mitkov. iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI'15, pages 74–76, Malaga, Spain, January 2015b.

Hernani Costa, Gloria Corpas Pastor, Miriam Seghiri, and Ruslan Mitkov. Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *New Horizons in Translation and Interpreting Studies (Full papers)*, pages 133–141, Geneva, Switzerland, December 2015c. Tradulex.

Hoang Cuong, Khalil Sima'an, and Ivan Titov. Adapting to all domains at once: Rewarding domain invariance in smt. *Transactions of the Association for Computational Linguistics*, 4:99–112, 2016. ISSN 2307-387X. URL https://www.transacl.org/ojs/index.php/tacl/article/view/768.

Joachim Daiber and Khalil Sima'an. Delimiting morphosyntactic search space with source-side reordering models. In *1st Deep Machine Translation Workshop*, pages 29–38, 2015a.

Joachim Daiber and Khalil Sima'an. Machine translation with source-predicted target morphology. In *Proceedings of the 15th Machine Translation Summit (MT Summit 2015)*, pages 283–296, Miami, USA, 2015b.

Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. Splitting compounds by semantic analogy. In *1st Deep Machine Translation Workshop*, pages 20–28. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, 2015.

Joachim Daiber, Miloš Stanojević, Wilker Aziz, and Khalil Sima'an. Examining the relationship between preordering and word order freedom in machine translation. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany, August 2016. Association for Computational Linguistics.

Asif Ekbal and Sivaji Bandyopadhyay. Voted ner system using appropriate unlabeled data. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, NEWS '09, pages 202–210, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-57-2. URL http://dl.acm.org/citation.cfm?id=1699705.1699749.

Asif Ekbal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. A modified joint source-channel model for transliteration. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL

'06, pages 191–198, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1273073.1273098.

Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer, and Michaela Regneri. Seedling: Building and using a seed corpus for the human language project. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 77–85, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W14/W14-2211.

Miquel Esplà-Gomis and Mikel L. Forcada. 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, August 2009.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL http://cs.jhu.edu/~ccb/publications/ppdb.pdf.

Rohit Gupta and Constantin Orăsan. Incorporating paraphrasing in translation memory matching and retrieval. In *Proceedings of the European Association of Machine Translation (EAMT-2014)*, pages 3–10, 2014.

Rohit Gupta, Hanna Bechara, and Constantin Orăsan. Intelligent translation memory matching and retrieval metric exploiting linguistic technology. In *Proceedings of the Translating and Computer 36*, pages 86–89, 2014.

Rohit Gupta, Constantin Orăsan, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. Can translation memories afford not to use paraphrasing? In *Proceedings of the 2015 Conference on European Association of Machine Translation (EAMT-2015)*, Antalya, Turkey, 2015.

Yifan He, Yanjun Ma, Josef van Genabith, and Andy Way. Bridging smt and tm with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL'10, pages 622–630, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Cuong Hoang and Khalil Sima'an. Latent domain translation models in mix-of-domains haystack. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1928–1939, 2014. URL http://aclweb.org/anthology/C/C14/C14-1182.pdf.

Cuong Hoang and Khalil Sima'an. Latent domain word alignment for heterogeneous corpora. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 398–408, 2015. URL http://aclweb.org/anthology/N/N15/N15-1043.pdf.

Chris Hokamp and Piyush Arora. DCU-SEManiacs at SemEval-2016 Task 1: Synthetic Paragram Embeddings for Semantic Textual Similarity. In *Proceedings of SemEval-2016*, pages 656–662, San Diego, California, USA, jun 2016. Association for Computational Linguistics.

Philipp Koehn and Jean Senellart. Convergence of Translation Memory and Statistical Machine Translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31, Denver, 2010.

Uri Lerner and Slav Petrov. Source-side classifier preordering for machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*, 2013.

Liangyou Li, Andy Way, and Qun Liu. A Discriminative Framework of Integrating Translation Memory Features into SMT. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas, Vol. 1: MT Researchers Track*, pages 249–260, Vancouver, BC, Canada, October 2014.

Liangyou Li, Carla Parra Escartín, and Qun Liu. Combining Translation Memories and Syntax-Based SMT: Experiments with real industrial data. *Baltic Journal of Modern Computing*, 4(2):165—-177, June 2016a.

Liangyou Li, Andy Way, and Qun Liu. Phrase-Level Combination of SMT and TM Using Constrained Word Lattice. In *Proceedings of ACL-2016: The 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, August 2016b. Association for Computational Linguistics.

Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 104–111, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220835.1220849. URL http://dx.doi.org/10.3115/1220835.1220849.

Varvara Logacheva and Lucia Specia. Confidence-based active learning methods for machine translation. In *EACL Workshop on Humans and Computer-assisted Translation*, HaCaT, pages 78–83, Gothenburg, Sweden, 2014a. URL http://www.aclweb.org/anthology/W14-0312.

Varvara Logacheva and Lucia Specia. A quality-based active sample selection strategy for statistical machine translation. In *Ninth International Conference on Language Resources and Evaluation*, LREC, pages 2690–2695, Reykjavik, Iceland, 2014b. ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/658_Paper.pdf.

Varvara Logacheva and Lucia Specia. The role of artificially generated negative data for quality estimation of machine translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 51–58, Antalya, Turkey, May 2015a. URL http://aclweb.org/anthology/W15-4907.

Varvara Logacheva and Lucia Specia. Phrase-level quality estimation for machine translation. In *Conference on Empirical Methods in Natural Language Processing (IWSLT)*, page 143–150, Beijing, China, July 2015b. Vietnam.

Varvara Logacheva, Chris Hokamp, and Lucia Specia. Data enhancement and selection strategies for the word-level quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 330–335, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-3039.

Vito Mandorino. 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.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N13-1090.

Nitesh Nahata, Tapas Nayak, Santanu Pal, and Sudip Kumar Naskar. 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.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, March 2003. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120103321337421.

Constantin Orăsan and Rohit Gupta, editors. *Proceedings of the First Workshop on Natural Language Processing for Translation Memories (NLP4TM-2015)*, RANLP 2015, Hissar, Bulgaria, September 2015.

Constantin Orăsan, Carla Parra, Eduard Barbu, and Marcello Federico, editors. *Proceedings of the 2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, LREC 2016, Portorož, Slovenia, may 2016.

Santanu Pal and Sudip Kumar Naskar. *Hybrid Approaches to Machine Translation*, chapter Hybrid Word Alignment, pages 57–75. Springer, 2016.

Santanu Pal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. Word alignment-based reordering of source chunks in pb-smt. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), 2014a.

Santanu Pal, Partha Pakray, and Sudip Kumar Naskar. Automatic building and using parallel resources for smt from comparable corpora. In *Hybrid Approaches to Translation (HyTra-2014) Workshop in 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, Gothenburg, Sweden, April 2014b.

Santanu Pal, Ankit Srivastava, Sandipan Dandapat, Josef van Genabith, and Andy Way. Usaar-dcu hybrid machine translation system for icon 2014. In *Proceedings of the 11th International Conference on Natural Language Processing (ICON-2014)*, Goa, India, 2014c.

Santanu Pal, Sudip Naskar, and Josef van Genabith. Uds-sant: English–german hybrid machine translation system. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 152–157, Lisbon, Portugal, September 2015a. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-3017.

Santanu Pal, Partha Pakray, Alexander Gelbukh, and Josef van Genabith. *Mining Parallel Resources for Machine Translation fromComparable Corpora*, pages 534–544. Springer International Publishing, Cham, 2015b. ISBN 978-3-319-18111-0. doi: 10.1007/978-3-319-18111-0_40. URL http://dx.doi.org/10.1007/978-3-319-18111-0_40.

Santanu Pal, Mihaela Vela, Sudip Kumar Naskar, and Josef van Genabith. Usaar-sape: An english–spanish statistical automatic post-editing system. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 216–221, Lisbon, Portugal, September 2015c. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-3026.

Santanu Pal, Sudip Kumar Naskar, and Josef van Genabith. Forest to string based statistical machine translation with hybrid word alignments. In *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science (CICLING-2016)*, Konya, Turkey, 2016a.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. A neural network based approach to automatic post-editing. In *Proceedings of the The 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin (Germany), August 2016b. ACL.

Santanu Pal, Marcos Zampieri, and Josef van Genabith. Usaar: An operation sequential model for automatic statistical post-editing. In *Proceedings of the ACL 2016 First Conference on Machine Translation (WMT16)*, Berlin, Germany, 11–12 August 2016c. ACL.

Vassilis Papavassiliou, Prokopis Prokopidis, and Gregor Thurmair. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2506.

Carla Parra Escartín. Creation of new tm segments: Fulfilling translators' wishes. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 1–8, Hissar, Bulgaria, September 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-5201.

Carla Parra Escartín and Lianet Sepúlveda Torres. Deliverable D6.5. Final User Evaluation. EXPERT Project, 2016a.

Carla Parra Escartín and Lianet Sepúlveda Torres. Deliverable D6.4. Improved Hybrid Translation Tool. EXPERT Project, 2016b.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Achim Ruopp and Fei Xia. Finding parallell texts on the web using cross-language information retrieval. In *Proceedings of the 2nd International Workshop on Cross Lingual Information Access*, pages 18–25, 2008.

Carolina Scarton and Lucia Specia. Document-level translation quality estimation: exploring discourse and pseudo-references. In *EAMT 2014*, pages 101–108, Dubrovnik, Croatia, 2014a.

Carolina Scarton and Lucia Specia. Exploring consensus in machine translation for quality estimation. In *Ninth Workshop on Statistical Machine Translation*, WMT, pages 342–347, Baltimore, Maryland, 2014b. URL http://www.aclweb.org/anthology/W14-3343.

Carolina Scarton and Lucia Specia. A quantitative analysis of discourse phenomena in machine translation. *Discours*, 16, 2015. doi: 10.4000/discours.9047. URL http://discours.revues.org/9047.

Carolina Scarton, Liling Tan, and Lucia Specia. Ushef and usaar-ushef participation in the wmt15 qe shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 336–341, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-3040.

Carolina Scarton, Daniel Beck, Kashif Shah, Karin Sim Smith, and Lucia Specia. Word embeddings and discourse information for Machine Translation Quality Estimation. In *Proceedings of the 11th Workshop on Statistical Machine Translation*, pages 2–7, Berlin, Germany, August 2016.

Kashif Shah, Varvara Logacheva, Gustavo Paetzold, Frédéric Blain, Daniel Beck, Fethi Bougares, and Lucia Specia. Shef-nn: Translation quality estimation with neural networks. In *Tenth Workshop on Statistical Machine Translation*, pages 338–343, Lisboa, Portugal, 2015. URL http://aclweb.org/anthology/W15-3041.

Radu Soricut and Franz Och. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N15-1186.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. Estimating the Sentence-Level Quality of Machine Translation Systems. In *13th Annual Meeting of the European Association for Machine Translation (EAMT-2009)*, pages 28–35, 2009.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. Quest - a translation quality estimation framework. In *51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL, pages 79–84, Sofia, Bulgaria, 2013. URL http://www.aclweb.org/anthology/P13-4014.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China, July

2015. Association for Computational Linguistics and The Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/P15-4020.

Liling Tan and Santanu Pal. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 201–206, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-3323.

Liling Tan, Rohit Gupta, and Josef van Genabith. Usaar-wlv: Hypernym generation with deep neural nets. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 932–937, Denver, Colorado, June 2015a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S15-2155.

Liling Tan, Josef van Genabith, and Francis Bond. Passive and pervasive use of bilingual dictionary in statistical machine translation. In *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 30–34, Beijing, China, July 2015b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-4105.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073478. URL http://dx.doi.org/10.3115/1073445.1073478.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596, 2005.

Kun Wang, Chengqing Zong, and Keh-Yih Su. Integrating Translation Memory into Phrase-Based Machine Translation during Decoding. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–21, Sofia, Bulgaria, August 2013.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. In *CoRR, abs/1511.08198*, 2015.

Friedel Wolff. 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.

Alena Zwahlen, Olivier Carnal, and Samuel Läubli. Automatic TM Cleaning through MT and POS Tagging: Autodesk's 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.