Exploiting data-driven hybrid approaches to translation in the EXPERT project

Constantin Orăsan[1], Carla Parra Escartín[2], Lianet Sepúlveda Torres[3], Eduard Barbu[4]

[1]University of Wolverhampton, United Kingdom;
[2]ADAPT Centre, Dublin City University, Ireland;
[3]Pangeanic, Spain
[4]University of Tartu, Estonia

C.Orasan@wlv.ac.uk, carla.parra@adaptcentre.ie, lisepul@gmail.com, eduard.barbu@ut.ee

August 2, 2018

Abstract

Researchers working in machine translation have benefited from the availability of large-scale corpora, and as a result in recent years an increasing number of empirical methods have been proposed. This chapter presents a brief overview of EXPERT (EXPloiting Empirical appRoaches to Translation), an FP7 EC-funded project whose main aim was to promote the research, development and use of data-driven hybrid language translation technology. Given the importance of translation memories in the everyday activities of professional translators, the chapter presents three research directions pursued in EXPERT which aimed to develop data-driven tools that are directly useful for translators.

11.1 Introduction

Technologies have transformed the way we work and this is also applicable to the translation industry. In the past 30-35 years, professional translators have experienced an increased technification of their work. Barely 30 years ago, a professional translator would not receive a translation assignment attached to an e-mail or via an FTP and yet, for the younger generation of professional translators receiving an assignment by electronic means is the only reality they know. In addition, as pointed out in several works such as Folaron [2010] and Kenny [2011], professional translators now have a myriad of tools available to use in the translation process.

All parties in the translation industry agree that Computer-Assisted Translation (CAT) tools are now their main working tool. Back in the early 1990s, when such tools started to be developed and used, they were little more than a set of Microsoft Word macros that integrated a Translation Memory (TM) engine and some sort of terminology management. Currently, these tools comprise of a wide variety of features that range from TM systems and terminology management tools, to Machine Translation (MT) plug-ins and Quality Assurance tools, with new features added on a regular basis. In addition to supporting translators during the translation process, CAT tools allow translators and project managers to carry out a complete translation cycle if needed.

Most of the components of a CAT tool rely on some kind of data in order to be useful to translators: translation memories rely on a database of previous translations, terminology management tools require access to term databases, whilst concordancers need corpora to extract examples of usages. For this reason, a significant amount of research in the field of translation technology has focused on the development of methods which can create such resources.

The EXPERT (EXPloiting Empirical appRoaches to Translation) project was an EC-funded FP7 project whose main aim was to promote the research, development and use of data-driven hybrid language translation technology.[1] The project appointed twelve Early Stage Researchers and three Experienced Researchers who worked on independent, but related, projects on various topics related to translation technology. The core objective of these projects was to create hybrid technologies which incorporate the best features of the existing corpus-based approaches and to improve the state-of-the-art of data-driven empirical methods used in translation.

The translation industry is now facing more challenges than ever. Translators are required to deliver high-quality professional translations, while having lower rates and increased time pressure imposed, as clients expect to get the translations they demand as fast as possible and for the lowest possible rate. One of the aims of the EXPERT project was to help translators with this issue by developing data driven translation technologies which speed the translation process up, whilst maintaining the quality of translation. In addition, the EXPERT project aimed to bridge the gap between academia and industry by applying some of the methods developed in the project to real life situations.

Prior to the EXPERT project, 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. This chapter offers an overview of several technologies developed in the EXPERT project which implemented the EXPERT solution. Given the limited space available and the fact that the EXPERT project focused on a variety

[1] http://expert-itn.eu

of topics, in most of the cases the research is presented only briefly with references to articles which provide further information.

The remainder of the chapter is structured as follows: Section 11.2 provides an overview of the EXPERT project, highlighting the most important outputs of the project, with emphasis on the hybrid data-driven research. Given the importance of translation memories in the work of professional translators, Section 11.3 presents the work on this topic carried out in the EXPERT project. The chapter finishes with conclusions.

11.2 The EXPERT project

The EXPERT (EXPloiting Empirical appRoaches to Translation) project was an EC-funded FP7 project under the People's programme. The main aim of the project was to train young researchers, namely Early Stage Researchers (ESRs) and Experienced Researchers (ERs), to promote the research, development and use of data-driven hybrid language translation technologies, and create future world leaders in the field. From the research perspective, the main objectives of the EXPERT project were to improve the corpus-based TM and MT technologies by addressing their shortcomings, and to create hybrid technologies which incorporate the best features of corpus-based approaches. The research also focused on how to consider the user requirements and translators' feedback in the translation process, as well as how to integrate linguistic knowledge that is usually ignored by the existing technologies.

The scientific work was organised into 15 individual projects, each linked to one of the main themes of the project: the user perspective, data collection and preparation, incorporation of language technology in translation memories, the human translator in the loop, and hybrid approaches to translation. This section presents a brief overview of the main research themes pursued in the project. A more detailed description can be found in [Orăsan et al., 2015].

The researchers had access to a vibrant training programme, which consisted of four large training events that ran across the whole consortium and engaged all the fellows: (1) Scientific and technological training, (2) Complementary skills training, (3) Scientific and technological workshop and, (4) Business showcase. In addition, they were involved in intersectoral and transnational mobilities via secondments and short visits to industrial and academic partners. Each researcher received training from their hosting institutions and all ESRs were registered on doctoral programmes.

The project was coordinated by the University of Wolverhampton, UK and consists of 6 academic partners: University of Wolverhampton, UK; University of Málaga, Spain; University of Sheffield, UK; University of Saarlandes, Germany; University of Amsterdam, Netherlands, and Dublin City University, Ireland; three companies: Translated, Italy; Hermes, Spain and Pangeanic, Spain and four associated partners: eTrad, Argentina; Wordfast, France; Unbabel,

Portugal and DFKI, Germany.

The rest of this section presents the work carried out across the main themes of the project.

The user perspective

The research on the user perspective sought to better understand the needs of professional translators by carrying out a large survey about their views and requirements regarding various technologies and their current work practices [Zaretskaya et al., 2015, 2018]. The survey showed that from the various technologies available, professional translators mostly avail of translation memories on a regular basis, and that the adoption of different tools depends very much on the translators' background. Because current CAT tools perform many tasks in addition to simple retrieval of previously translated segments, the survey revealed that translators use these tools to perform a variety of other tasks such as terminology management and quality assurance.

Under the same theme of user perspective, Hokamp and Liu [2015] proposed HandyCAT, an open source CAT tool[2] that allows the user to easily add or remove graphical elements and data services to/from the interface. Moreover, new components can be directly plugged into the relevant part of the translation data model. These features make HandyCAT an ideal platform for developing prototypes and conducting user studies with new components.

Data collection and preparation

The focus of the project was on data-driven technologies. For this reason, extensive research on data-collection and preparation was also carried out. Costa et al. [2015] developed iCorpora, a tool which can semi-automatically compile monolingual and multilingual parallel and comparable corpora from the web. In addition to compiling corpora, the tool also enables users to manage corpora and exploit them. Barbu [2015] worked on the cleaning of translation memories. The task of cleaning translation memories became the focus of a shared task organised by the consortium and is presented in more detail in Section 11.3.3.

Language technology in translation memory

As demonstrated by the survey mentioned above, translation memories are among the most successfully used tools by professional translators. However, most of these tools rely on little language processing when they match and retrieve segments. Section 11.3 presents the research carried out in the EXPERT project that incorporates information from a paraphrase database into matching and retrieval from translation memories, and shows how this can improve the productivity of professional translators, and how to deal with large translation memories. In the same vein of research, Tan and Pal [2014] proposed several methods for terminology extraction

[2] http://handycat.github.io/

and ontology induction with the aim of integrating them in translation memories and statistical machine translation.

The human translator in the loop

The work dedicated to the "human translator in the loop" investigated approaches to inform end-users about the quality of translations, as well as learning from their feedback on the quality of translations to improve translation systems and workflows. The work focused on ways of collecting and extracting useful information from post-edited sentences to feedback into SMT systems [Logacheva and Specia, 2015], as well as discourse level quality estimation, a topic largely neglected by the research community [Scarton and Specia, 2014]. Carla Parra Escartín et al. [2017] analysed the work of professional translators when they are asked to post-edit segments of various qualities in an attempt to better understand how the quality of automatically translated segments influences the post-editing process.

To better understand the post-editing process, researchers working on the EXPERT project developed CATaLog[3] [Nayek et al., 2015, 2016] and CATaLog Online [Pal et al., 2016], the online version of CATaLog. Both tools are language independent CAT tools which provide a user-friendly CAT environment to post-edit translation memory segments and machine translation output. They were implemented to minimize the translators' and post-editors' efforts during the post-editing task. One of the main innovations of *CATaLog online* consists of integrating a colour coded scheme both for the source and target segments to highlight the chunks in a particular segment that should be changed. Whilst most CAT tools highlight fuzzy match differences, this is only done on the source side and it is left to the translator to locate the part of the target segment that needs to be changed. Additionally, the tool includes automatic logging of user activity. It automatically records keystrokes, cursor positions, text selection and mouse clicks together with the time spent post-editing each segment. This way it collects a wide range of logs with post-editors' feedback that can be very useful for research on post-editing and can also be used as training materials for Automatic Post-Editing tasks. These features make the tool ideal for MT developers and researchers in translation studies.

Hybrid approaches to translation

A significant amount of research was carried out on hybrid approaches to translation and delivered a general framework for the combination of SMT and TM which outperforms the state-of-the-art work [Li et al., 2016], better ways of incorporating a dependency tree into a statistical machine translation model [Li et al., 2015], methods for performing source-side pre-ordering for improving the quality of SMT [Daiber and Sima'an, 2015], and a method that produces better translations by considering the domain of the text to be translated [Cuong et al., 2016].

[3] https://github.com/santanupal1980/CATaLog

This section briefly presented the context in which the research described in this chapter took place. Given that the EXPERT project was a four year project, it produced much more than the research described here. For example, given the importance of semantic text similarity for many of the topics researched in the project, there were a number of systems submitted to the task of Semantic Text Similarity organised at SemEval conferences. The same happened with various evaluation metrics submitted at WMT workshops. The project's webpage provides the complete list of publications which resulted from the project and links to the resources released.

11.3 Translation Memories

As shown in Zaretskaya et al. [2018], Translation Memory systems are very important tools for professional translators and constitute a key component of CAT tools. Translation Memories store past translations which can be retrieved when either an identical or a very similar new sentence has to be translated. This process is called *TM leveraging*. In order to leverage past translations, TM systems rely on an algorithm to measure the similarity between a sentence to be translated and those stored in the TM. For each new sentence, the TM system computes this similarity and assigns all identical or similar translations a score called the Fuzzy Match Score (FMS). Given that the translation memories store the sentence in both the source language and its translation, TM systems will offer the translation of the sentence from the translation memory with the highest FMS as a suggested translation for the new sentence. This is done even in the cases where the two sentences are not identical, the assumption being that if they are similar enough the effort to edit the suggested translation is lower than the effort necessary to translate from scratch. In fact, to enhance the editing process, the TM tool additionally highlights the differences between the new sentence to be translated and the one stored in the TM thus allowing the translator to quickly identify parts that need to be edited.

Translators are used to edit the so-called fuzzy matches and the TM leveraging is also used to compute rates and allocate resources at the planning stage of a translation project. As the FMS decreases, sentences are more difficult to edit and at some point they are not worth editing. That is why in the translation industry a threshold of 75% FMS is used. Segments getting a 75% FMS or higher undergo fuzzy match editing, and segments below that threshold are translated from scratch.

The EXPERT project proposed two main ways of improving TM systems. The first focused on improving the way in which TM systems carry out the TM leveraging process, by proposing new ways of identifying similar sentences in the translation memories. Section 11.3.1 proposes a new method for calculating the similarity between sentences which goes beyond surface matching and incorporates a database of paraphrases in the process. The usefulness of translation memory systems improves when translators have access to larger

translation memories. Section 11.3.2 describes a fast and scalable tool for translation memory management. Another direction of research investigated in the EXPERT project focused on the task of curating TMs to ensure their high quality, and automatically cleaning them. This is presented in Section 11.3.3.

11.3.1 Incorporating semantic information in the matching and retrieval process

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 a such segment in the available TMs. As previously explained, TMs will not only retrieve the exact matches found, but also fuzzy matches (i.e. similar segments to the one that needs to be newly translated). Fuzzy matches are retrieved using some sort of Edit Distance Metric such as Levenshtein Distance.

Gupta and Orăsan [2014] explore the integration of paraphrases in matching and retrieval from TMs using Edit Distance in an approach based on greedy approximation and dynamic programming. The proposed method modifies Levenshtein Distance to take into consideration paraphrases extracted from PPDB [Ganitkevitch et al., 2013] when it is calculated. In addition, it is possible to paraphrase existing TMs to allow for offline processing of data and alleviate the need for translators to install additional software. Their system is based on the following 5-step pipeline:

1. Read the TMs.
2. Collect all paraphrases from the paraphrase database and classify them in classes:
   (a) Paraphrases involving one word on both the source and target side.
   (b) Paraphrases involving multiple words on both sides but differing in one word only.
   (c) Paraphrases involving multiple words but the same number of words on both sides.
   (d) Paraphrases with differing number of words on the source and target sides.
3. Store all the paraphrases for each segment in the TM.
4. Read the file to be translated.
5. Get all paraphrases for all segments in the file to be translated, classify them and retrieve the most similar segment above a predefined threshold.

They report a significant improvement in both retrieval and translation of the retrieved segments. This research was further expanded with a human centered evaluation in which the quality of semantically informed TM fuzzy matches were assessed based on editing time or keystrokes [Gupta et al., 2015]. This evaluation revealed

that both the editing time and the number of keystrokes are reduced when the enhanced edit distance metric is used, without a decrease in the quality of translation. The tool has been publicly released under an Apache License 2.0 and is available on GitHub[4].

11.3.2 ActivaTM a Translation Memory Management (TMM) system

The previous section demonstrated how it is possible to improve the matching and retrieval from translation memories by incorporating semantic information from a database of paraphrases in the matching algorithm. This can be very useful for professional translators, but is not enough. The survey carried out by Zaretskaya et al. [2018] showed that a fast response is an essential feature for translation memories. When working on large translation projects, translators usually have access to massive background translation memories, which are sometimes augmented with input from fully automatic translation engines. In these cases speed of access can become a problem and specialist solutions have to be sought.

ActivaTM is a fast and scalable Translation Memory Management (TMM) system developed in the EXPERT project to ensure fast access to large translation memories. It is based on a full-text search engine which has the ultimate goal of providing Translation Memory capabilities for a hybrid machine translation workflow. It can be used in a CAT environment to provide almost perfect translations to the human user with markups highlighting the translated segments that need to be checked manually for correctness.

This TMM system was designed in such a way that it can be successfully integrated into an online CAT tool environment, where several translators work simultaneously in the same project, adding and updating TM entries. *ActivaTM* can also outperform pure Statistical Machine Translation (SMT) when a good TM match is found and in the task of automatic website translation. Preliminary experiments showed that the *ActivaTM* system overcomes the limitations of current TM systems in terms of storage and concordance searches. The remainder of this section presents the main features of ActivaTM.

11.3.2.1 ActivaTM design and Capabilities

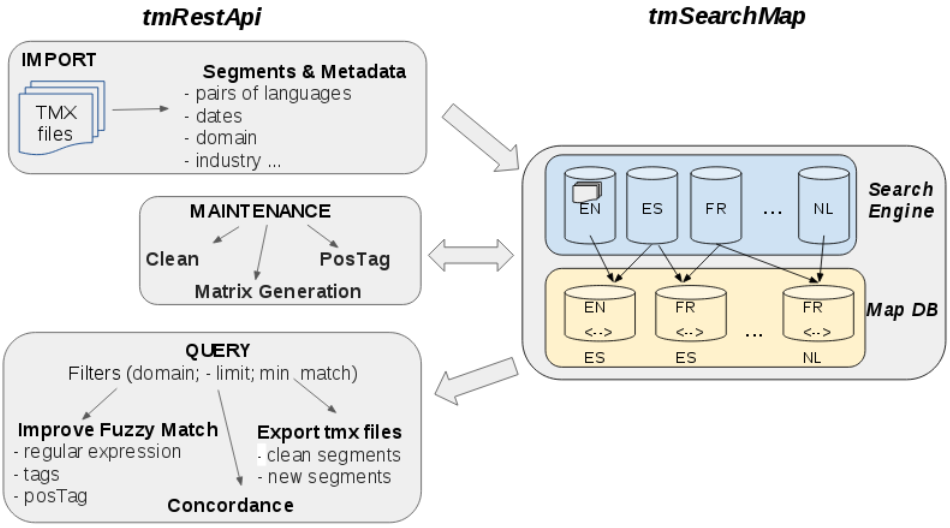
Figure 11.1 presents an overview of the ActivaTM system. It consists of two principal components **tmSearchMap** and **tmRestAPI** with the aim to ensure the following requirements:

- **Great storage capacity**: The system has the capacity to store large numbers of segments (over 10M), along with their corresponding metadata (source and target language, segment creation and modification date, part-of-speech tags for all tokens in a given segment and for both languages, domains, etc.).

[4] https://github.com/rohitguptacs/TMAdvanced

- **Fast and efficient retrieval algorithm**: The system is able to retrieve fuzzy matches quickly regardless of their FMS score.
- **Reasonable import time of new segments**: It is foreseen that occasionally the existing TMs will need to be updated by importing a massive number of new segments in the TMX format or similar format. ActivaTM is able to achieve this task within a reasonable time.
- **Effective segment filtering, retrieval and export**: The system is able to retrieve sets of segments fulfilling certain criteria (e.g. domain, date, time span, file name, terms appearing in the source or target language, etc.). Such subsets can be exported as one or several TMs in the TMX format that can subsequently be used to train MT systems, like those available in the PangeaMT platform[5].

Figure 11.1: Main components of ActivaTM.



11.3.2.2 tmSearchMap

As the name suggests, *tmSearchMap* is the component in charge of retrieving segments from the translation memory. It is based on Elasticsearch[6] and takes advantage of its Information Retrieval based indexing technique to speed up the time-consuming TM retrieval procedure. Elasticsearch was selected because is a mature project that dominates the open-source search engine market, and supports fast mapping of source segments considering exact match, fuzzy

[5] http://pangeamt.com/en
[6] https://www.elastic.co/

match and regular expression. tmSearchMap consists of two principal applications: Search Engine and MapDB.

The purpose of the *Search Engine* is to store monolingual indices of segments and provide a flexible search interface, whilst *MapDB* aims to complement Search Engine by storing pairs of bidirectional mappings. MapDB stores both the source and target texts together with their metadata, extracted from TMX[7] files, such as domain, industry, type, organisation, several dates etc. MapDB also supports quick bulk import and update operations. The purpose of the update operation is to enable future updates of a segment including updating the modification date to indicate when translators edited a segment. This design enables the corresponding id and text of the target language segment to be quickly retrieved, after identifying a match in a monolingual index. Additionally, the design saves a significant amount of memory by only storing each unique segment once, which is necessary when dealing with large translation memories.

Using the above design, a query to ActivaTM is conducted as follows, taking the EN-ES language pair as an example:

- A client queries ActivaTM by providing a source (EN) language segment.
- ActivaTM uses its search engine to identify the most suitable segment in the EN index and retrieves its id.
- MapDB index EN-ES is queried using the retrieved id and then the bilingual properties are retrieved, returning the stored translation to the client.

11.3.2.3 tmRestAPI

The *tmRestAPI* implements a series of operations which allow importing, maintenance, and query of TMs.

Importing

From time to time, it is foreseen that the existing TMs will need to be updated by importing new segments in TMX formats. tmRestApi is capable of importing large numbers of segments (over 10M), along with their corresponding metadata (source and target language, segment creation and modification date, domains, industry, etc). The ActivaTM system implements a TMX parser to extract the above properties from the input files and is able to store the new pairs of segments in a database within a reasonable time.

Maintenance

Maintenance tasks aim to improve the quality of the existing data, generate new data, and aggregate new properties to the existing data. To increase efficiency and minimise interference with the work done

[7]TMX stands for Translation Memory eXchange and is an XML based format for exchanging translation memories between computers. The details of the standard can be accessed at https://www.gala-global.org/tmx-14b

by translators, all of these processes occur in the background and are performed on specific segments which are selected on the basis of a predefined set of characteristics. The following tasks are performed during the maintenance: part-of-speech (POS) tagging, cleaning and matrix generation.

Part-of-Speech tagging: In order to improve the retrieval operation, all the tokens in source and target segments are tagged with part-of-speech information. ActivaTM is able to use different taggers, depending on the language to be analysed, their precision and performance. For example, it uses *TreeTagger*[8] [Schmid, 1994] for segments in English, Spanish and French, for Japanese it employs *KyTea*[9] [Neubig et al., 2011] whereas for other languages it relies on *RDRPOSTagger*[10] [Nguyen et al., 2014], which includes the pre-trained Universal POS tagging models for 40 languages. To allow for comparisons across languages, the Universal PoS tagset [Petrov et al., 2012] is also used.

Cleaning: As discussed in the next section, it is not unusual to have noise in TM files. The cleaning task aims to identify and penalise pairs of noisy segments on a database. This will ensure that during the query, Elasticsearch does not rank spurious segments among the best. Currently, ActivaTM distinguishes most punctuation and numerical inconsistencies in the source language and in the target language.

Matrix generation: This task implements a triangulation algorithm which takes advantage of the tmSearchEngine design to create new pairs of segments from existing segments. The algorithm considers the stored data as an undirected graph where each monolingual segment is a node and each bilingual entry is an arc connecting nodes of different languages which are known to be correct. In this way, if for one of the segments we have translations into more than one target language, we can generate translations between these target languages even if they are not explicitly specified.

tmRestApi: Query Task

ActivaTM takes advantage of the Elasticsearch powerful query language to implement a fast and efficient retrieval algorithm. The sets of segments retrieved using this language can be restricted to fulfil certain criteria such as coming from a specific domain, containing certain terms in the source or target language, and/or having specific time stamps. These sets can be exported as one or several TMs in a TMX format and used to train customised translation engines (both TM and SMT engines).

An innovation of ActivaTM is its fuzzy match score, which was created specifically for the tool and leads to better ranking of the segments retrieved by Elasticsearch. The FMS is based on the well-known Levenshtein distance, however the score between the query

[8] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[9] http://www.phontron.com/kytea/
[10] http://rdrpostagger.sourceforge.net/

and the TM source is calculated taking into account the similarities of both strings considering the following features: characters, words, punctuation and stop words.

In addition, ActivaTM exploits existing linguistic knowledge to improve the fuzzy matching algorithm, and as a consequence the TM leveraging. The fuzzy match algorithm consists of a pipeline, that integrates several language dependent and independent features, such as: regular expressions, tag processing and part-of-speech sequence matching. Regular expressions are used to improve the recognition of placeable and localisable elements (e.g. numbers, URLs, etc.). Part-of-speech matches are used to detect grammatical similarities between source and target segments. Currently, only segments having a very similar grammatical structure benefit from part-of-speech matches. This procedure is also used to identify mismatches between source and target segments, and relies on a glossary or the output of a SMT system to obtain translations of the mismatched words in the target segment.

11.3.3 Translation Memory Cleaning

As reiterated throughout this chapter, translation memories are very important resources for translators. However, in order to be beneficial to translators, the contents already stored in a TM must be of high quality and correct. This is not always the case when TMs are built by communities[11] or they are automatically harvested from the web. Manual cleaning is expensive and sometimes not possible due to the lack of domain experts. For this reason, the researchers involved in the EXPERT project proposed a method for the automatic cleaning of translation memories and organised a shared task which focused on cleaning of translation memories.

11.3.3.1 TM cleaner

Barbu [2015] has developed a machine learning based tool which is able to identify false translations in pairs of segments stored in translation memories. The system is trained and tested on a dataset extracted from MyMemory. Analysis of the data revealed four main sources of errors:

- random text where a contributor copies random text for the source and/or target segment. These cases usually indicate a malevolent contributor.
- chat-like contributions when the translation memory users exchange messages instead of providing translations. For example the English text "How are you?" translates in Italian as "Come

[11] For example, MyMemory (`https://mymemory.translated.net/`) [Trombetti, 2009] allows anyone to register and translate using their online portal. During this process, users also build translation memories which can be used in other translation processes. Because anyone can participate, by default there is no quality control on the translation memories and some of the entries contain mistakes.

stai?". Instead of providing this translation the contributor answers "Bene" ("Fine").

- language errors when languages of the source or target segments are mistaken or swapped. This type of error usually occurs in TMs which have several target languages.
- partial translations where only part of the source segment is translated.

The method proposed uses 17 features to train a machine learning classifier to identify false translation pairs. These features cover a range of phenomena which can indicate correct or incorrect translations such as presence of URLs, tags or email addressees, the cosine similarities between the use of punctuation, tags, email addresses, and URLs between the source and the target. The full list of features and their descriptions can be found in Barbu [2015].

Evaluation of six different machine learning algorithms revealed that with the exception of Naive Bayes, all of them perform much better than two baselines (a random baseline and a random baseline which respects the training set class distribution). However, a detailed error analysis of the results shows that the classifiers produce a large number of false negatives, which means that around 10% of good examples would need to be discarded. A solution to this problem is to develop methods which have higher precision, even if this means a lower recall.

An enhanced version of the TM cleaner is freely available on GitHub[12]. The main differences between the algorithm presented in Barbu [2015] and the implementation on GitHub are features to make it easily usable by the translation industry. The main differences are:

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 with the training model.
2. Integration of the Fastalign word aligner as a web service: Like above, this component is meant to replace the automatic translation. Based on the alignments returned by the word aligner, new features are computed (e.g. number of aligned words in source and target segments). For more details please see [Barbu, 2017]
3. Addition of two operating modes: 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.
4. Passing arguments through the command line: It is now possible, to indicate the machine-learning algorithm that will be used for classification.

[12] https://github.com/SoimulPatriei/TMCleaner

5. Implementation of hand written rules for keeping/deleting certain bilingual segments: These hand written rules are necessary to decide in certain 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.
6. 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 websites and TMs. Moreover, the final version of the tool has been implemented on an iterative process based on annotating the data and evaluating it using the evaluation module. This iterative process has been followed to boost the performance of the cleaner.

11.3.3.2 Automatic Translation Memory Cleaning Shared Task

This shared task was inspired by the work carried out by [Barbu, 2015] in the EXPERT project as presented above, and was one of the outcomes of the First Workshop on Natural Language Processing for Translation Memories (NLP4TM)[13]. The purpose of the first Automatic Translation Memory Cleaning Shared Task was to invite teams from both academia and industry to tackle the problem of cleaning TMs and submit their automatic systems for evaluation. As this was the first shared task on this topic, the focus was on learning to better define the task and on understanding what are the most promising approaches to tackle the problem. The proposed task consisted of identifying translation units that had to be discarded because they were inaccurate translations of each other, or corrected as they contained orthotypographical errors such as missing punctuation marks or misspellings.

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").

[13] http://rgcl.wlv.ac.uk/nlp4tm/

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

1. Binary Classification (I): In this task, it was only necessary 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").

The data was, for the most part, sampled from the public part of MyMemory. In the initial phase, we extracted approximately 30,000 translation units (TUs) for each language pair. The TUs were heterogeneous and belonged to different domains, ranging from medicine and physics to colloquial conversations. A set of filters was applied in order to reduce this number to 10,000 units per language, from which approximately 3000 TUs per language pair were manually selected. Since the proportion of units containing incorrect translations is low, to facilitate their manual selection we computed the cosine similarity score between the machine translation of the English segment and the target segment of the TU. The hypothesis to test was that low cosine similarity scores (less than 0.3) can signal bad translations. Finally, we ensured that the manually selected TUs did not contain inappropriate language or other errors that could not be identified automatically. The data was manually annotated by two native speakers.

In total six teams participated in the shared-task, by submitting a total of 45 runs. Barbu et al. [2016] contains a detailed description of the participating teams and a comparative evaluation of their results. In addition, the reports from each of the participating teams can be found on the shared-task's webpage[14].

11.4 Conclusion

This chapter presented the main research topics addressed in the EXPERT project and summarized some of the innovations of

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

the EXPERT project related to data-driven hybrid approaches to translation. Some of the researchers employed in EXPERT focused on improving already existing algorithms with linguistic information, whilst others have researched how to create new tools that can be used in the translation industry. As a result, the *TMAdvanced* tool developed by Gupta and Orăsan [2014] can already be used by any translator or translation company, as can the *ActivaTM*[15] and the *TM Cleaner*.

The CAT tools CATaLog [Nayek et al., 2015, Pal et al., 2016] and HandyCAT [Hokamp and Liu, 2015], and the terminology management system proposed by Hokamp [2015], are also examples of how academic research can produce Open Source tools that aim to fulfill all the features of existing CAT tools and whilst adding new functionalities with the sole purpose of helping translators translate better and focus on the task at hand: delivering high quality translations in a timely manner.

Several EXPERT researchers have explored ways of integrating new advances in Computational Linguistics and Machine Translation in the translation workflow. The research carried out in the EXPERT project proved there is room for a successful hybridization of the translation workflow and such hybridization may be implemented in different components with a unique goal: enabling the end users (i.e. the translators) to work more efficiently and effectively as a benefit of the research undertaken.

Acknowledgment

We would like to acknowledge the contribution of all the partners and all the researchers to the project. This chapter would not have been possible without their contribution. The research described here was partially 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. Carla Parra Escartín is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 713567, and Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106).

References

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.

[15]Due to the fact that ActivaTM was designed to run part of a proprietary it is not available as a stand-alone tool and currently cannot be accessed for free.

Eduard Barbu. Ensembles of classifiers for cleaning web parallel corpora and translation memories. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 71–77, Varna, Bulgaria, September 2017. URL https://doi.org/10.26615/978-954-452-049-6_011.

Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Constantin Orăsan, and Marcello Federico. The first Automatic Translation Memory Cleaning Shared Task. *Machine Translation*, 30(3-4):145–166, dec 2016. ISSN 0922-6567. doi: 10.1007/s10590-016-9183-x. URL http://link.springer.com/10.1007/s10590-016-9183-x.

Carla Parra Escartín, Hanna Béchara, and Constantin Orăsan. Questing for Quality Estimation A User Study. *The Prague Bulletin of Mathematical Linguistics*, 108:343–354, 2017. doi: 10.1515/pralin-2017-0032. URL https://ufal.mff.cuni.cz/pbml/108/art-bechara-escartin-orasan.pdf.

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, Genebra, Switzerland, December 2015. 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. URL https://transacl.org/ojs/index.php/tacl/article/view/768.

Joachim Daiber and Khalil Sima'an. Machine Translation with Source-Predicted Target Morphology. In *Proceedings of MT Summit XV*, Miami, Florida, 2015.

Deborah Folaron. Translation tools. In *Handbook of Translation Studies*, volume 1, pages 429–436. John Benjamins Publishing Co., Amsterdam; Philadelphia, 2010.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In *Proceedings of NAACL-HLT 2013*, pages 758–764, Atlanta, Georgia, jun 2013. URL http://www.aclweb.org/anthology/N13-1092.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, 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.

Chris Hokamp. Leveraging NLP technologies and linked open data to create better CAT tools. *Localisation Focus - The International Journal of Localisation*, (14), 2015.

Chris Hokamp and Qun Liu. Handycat: The Flexible CAT Tool for Translation Research. In *Demo presented at EAMT 2015*, pages 15–19, Istanbul, Turkey, May 2015.

Dorothy Kenny. Electronic Tools and Resources for Translators. In Kirsten Malmkjær and Kevin Windle, editors, *The Oxford Handbook of Translation Studies*, pages 455–472. Oxford University Press, Oxford, England, 2011. URL http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199239306.001.0001/oxfordhb-9780199239306-e-031.

Liangyou Li, Andy Way, and Qun Liu. Dependency Graph-to-String Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 33–43, Lisbon, Portugal, 2015. ISBN 9781941643327. URL http://aclweb.org/anthology/D15-1004.

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 2016.

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

Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. CATaLog: New approaches to TM and post editing interfaces. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 36–42, Hissar, Bulgaria, September 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-5206.

Tapas Nayek, Santanu Pal, Sudip Kumar Naskar, Sivaji Bandyopadhyay, and Josef van Genabith. Beyond translation memories: Generating translation suggestions based on parsing and pos tagging. In *Proceedings of the 2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, Portorož, Slovenia, May 2016.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 529–533, Portland, Oregon, USA, June 2011. URL http://www.phontron.com/paper/neubig11aclshort.pdf.

Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. RDRPOSTagger: A ripple down rules-based part-of-speech tagger, 2014.

Constantin Orăsan, Alessandro Cattelan, Gloria Corpas Pastor, Josef van Genabith, Manuel Herranz, Juan José Arevalillo, Qun Liu, Khalil Sima'an, and Lucia Specia. The EXPERT Project: Advancing the State of the Art in Hybrid Translation Technologies. In *Proceedings of Translating and the Computer 37*, London, UK, 2015.

Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Tapas Nayak, Mihaela Vela, and Josef van Genabith. CATaLog online: Porting a post-editing tool to the web. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Carolina Scarton and Lucia Specia. Document-level translation quality estimation: exploring discourse and pseudo-references. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014)*, pages 101 – 108, Dubrovnik, Croatia, 2014.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees, 1994.

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, 2014.

Marco Trombetti. Creating the worlds largest translation memory. In *MT summit XII: proceedings of the twelfth machine translation summit*, pages 9 – 16, Ottawa, ON, Canada, 2009.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596, 2005.

Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Seghiri. Translators' requirements for translation technologies: Results of a user survey. In *Proceedings of the AIETI7 Conference. New Horizons is Translation and Interpreting Studies (AIETI)*, Málaga, Spain, 2015.

Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Seghiri. User Perspective on Translation Tools: Findings of a User Survey. In Gloria Corpas Pastor and Isabel Duran, editors, *Trends in E-tools and Resources for Translators and Interpreters*, pages 37 – 56. Brill, 2018.