Reciprocal Enrichment Between Basque Wikipedia and Machine Translation

Iñaki Alegria, Unai Cabezon, Unai Fernandez de Betoño, Gorka Labaka, Aingeru Mayor, Kepa Sarasola and Arkaitz Zubiaga

Abstract In this chapter, we define a collaboration framework that enables Wikipedia editors to generate new articles while they help development of Machine Translation (MT) systems by providing post-edition logs. This collaboration framework was tested with editors of Basque Wikipedia. Their post-editing of Computer Science articles has been used to improve the output of a Spanish to Basque MT system called Matxin. For the collaboration between editors and researchers, we selected a set of 100 articles from the Spanish Wikipedia. These articles would then be used as the source texts to be translated into Basque using the MT engine. A group of volunteers from Basque Wikipedia reviewed and corrected the raw MT translations. This collaboration ultimately produced two main benefits: (i) the change logs that would potentially help improve the MT engine by using an automated statistical post-editing system , and (ii) the growth of Basque Wikipedia. The results show that this process can improve the accuracy of an Rule Based MT (RBMT) system in nearly 10% benefiting from the post-edition of 50,000 words in the Computer

Iñaki Alegria
Ixa Group, University of the Basque Country UPV/EHU, e-mail: i.alegria@ehu.es

Unai Cabezon
Ixa Group, University of the Basque Country, e-mail: ucabezon001@ikasle.ehu.es

Unai Fernandez de Betoño
Basque Wikipedia and University of the Basque Country,
e-mail: unai.fernandezdebetono@ehu.es

Gorka Labaka
Ixa Group, University of the Basque Country, e-mail: gorka.labaka@ehu.es

Aingeru Mayor
Ixa Group, University of the Basque Country, e-mail: aingeru@ehu.es

Kepa Sarasola
Ixa Group, University of the Basque CountryU, e-mail: kepa.sarasola@ehu.es

Arkaitz Zubiaga
Basque Wikipedia and Queens College, CUNY, CS Department, Blender Lab, New York,
e-mail: arkaitz.zubiaga@gmail.com

Science domain. We believe that our conclusions can be extended to MT engines involving other less-resourced languages lacking large parallel corpora or frequently updated lexical knowledge, as well as to other domains.

1 Introduction

One of the key features on the success of Wikipedia, the popular and open online encyclopedia, is that it is available in more than 200 languages. This enables the availability of a large set of articles in different languages. The effort of Wikipedia editors to keep contents updated, however, increases as the language has a smaller community of editors. Because of this, less-resourced languages with smaller number of editors cannot keep pace with the rapid growth of top languages such as English Wikipedia. To reduce the impact of this, editors of small Wikipedias can take advantage of contents produced in top languages, so they can generate large amounts of information by translating those. To relax such process of translating large amounts of information, machine translation provides a partially automated solution to potentially facilitate article generation [13]. This presents the issue that current machine translation systems generate inaccurate translations that require substantial post-editing by human editors. We argue that creatively combining machine translation and human editing can benefit both article generation on Wikipedia, and the development of accurate machine translation systems.

In this chapter, we introduce our methodology to enable collaboration between Wikipedia editors and researchers, as well as the system we have developed accordingly. This system allows to generate new articles by editing machine translation outputs, while editors help improve a machine translation system. Specifically, editors of the Basque Wikipedia have used this system to collaborate with the University of the Basque Country (UPV/EHU) producing articles in Basque language while helping improve an existing Spanish-Basque machine translation (MT) system called Matxin [9]. We believe that amateur translators can benefit from MT rather than professional translators.

To perform such a collaboration between editors and researchers, a set of 100 articles were selected from Spanish Wikipedia to be translated into Basque using the MT engine. A group of volunteers from Basque Wikipedia reviewed and corrected these raw translations. In the correction process, they could either post-edit a text to fix errors, or retranslate it when the machine-provided translation was inaccurate. We logged their changes, and stored the final article generated. This process ultimately produced two main benefits: (i) the change logs potentially help improve the MT engine by using an automated statistical post-editor [11], and (ii) the generated articles help expand the Basque Wikipedia. The results show that this process can improve the accuracy of an Rule Based MT (RBMT) system in nearly 10% benefiting from the post-edition of 50,000 words in the Computer Science domain.

The remainder of the chapter is organized as follows: section 2 provides an overview of the most representative features of Basque language, as well as a sum-

mary of previous research on statistical post-edition (SPE), and collaborative work and MT; section 3 describes the methodology used to build the post-editing system; section 4 outlines and discusses the results and resources obtained through the collaborative work; finally, section 5 concludes the chapter and sketches our future research plans.

2 Background

In this section we briefly describe features of Basque language, and summarize previous research on collaborative work for machine translation and automated statistical post-edition.

2.1 Basque Language

Basque language presents particular characteristics, making it different from most European languages. This also makes translating into Basque a challenging task compared to other languages that share some sort of similarities. As an agglutinative language, many morpho-syntactic information that most European languages express in multiple words are expressed in a single word using suffixes in Basque. For instance, while Spanish and English use prepositions and articles, in Basque, suffixes are added to the last word of the noun-phrase; similarly, conjunctions are attached at the end of the verbal phrase.

Additionally, syntactic differences can also be found when looking into word orderings. These include: (i) modifiers of both verbs and noun-phrases are ordered differently in Basque and in Spanish; (ii) prepositional phrases attached to noun-phrases precede the noun phrase instead of following it; iii) having very flexible ordering of sentence constituents, a neutral ordering suggests placing the verb at the end of the sentence and after the subject, object and any additional verb modifiers.

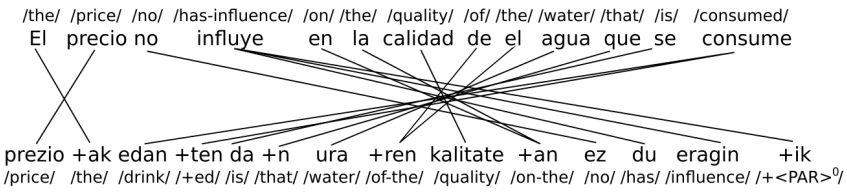

Fig. 1 Comparison of word alignment for a sentence in Spanish and Basque (their transcription to English is "The price does not affect the quality of the drinking water")

Figure 1 shows an example that compares word alignment for the Spanish sentence “El precio no influye en la calidad del agua que se consume” (The price does not affect the quality of the drinking water) and its Basque translation “Prezioak edaten den uraren kalitatean ez du eraginik”.

All those differences make translating from Spanish (or English) into Basque a challenging process that involves both morphological and syntactical features. On top of that, the fact that Basque is a low resourced language[1] makes the development of a MT system an even more ambitious undertaking.

2.2 Related Work on Collaboration initiatives and Machine Translation

Most MT engines make use of translations produced by humans. Specifically, translation repositories (usually referred to as translation memories,TMs) or parallel corpora are harnessed to learn translation models [13]. The use of public TMs has helped in the development and improvement of MT engines, and many companies have shared their memories to this end (e.g. TAUS[2]).

The chapter *“Building Multilingual Language Resources in Web Localisation: A Crowdsourcing Approach”* of this book describes a client-server architecture to share and use translation memories, which can be used to build (or improve) MT systems.

An alternative solution for improving MT engines is taking advantage of post-edition, i.e., the process of correcting MT outputs. The outcome of a post-editing process can be used in several ways:

- As a quality baseline to evaluate MT engines.
- As a resource that provides new TMs to help improve an MT engine.
- As a set of *automatic output/post-edited output* pairs that enables to learn an automatic post-editor (see Section 2.3). We use post-editing to this end on our system.

Popular MT engines include a post-edition interface to fix translations. For instance, Google Translate[3] allows its users to post-edit translations by replacing or reordering words. These corrections, which are only internally available to Google, provide valuable knowledge to enhance the system for future translations.

Asia Online is leading a project which aims to translate contents from English Wikipedia into Thai. In 2011, the company translated 3.5 million Wikipedia articles

[1] There are around 700,000 speakers, around 25% of the total population of the Basque Country, but the use of Basque in industry and especially in Information and Communication Technology is not widespread

[2] www.translationautomation.com

[3] http://translate.google.com

using MT and they are planning to improve them collaboratively.[4] Further details on the selected methodology are not available yet.

Other companies such as Lingotek,[5] sell *Collaborative Translation Platforms* that include post-edition capabilities.[6]

For our collaborative work, we use OmegaT,[7] an open source Computer Aided Translation (CAT) tool.

2.3 Related Work on Training a Post-Editing System

Statistical post-editing , as described by Simard et al. [11], is the process of training a Statistical Machine Translation (SMT) system to translate from rule-based MT (RBMT) outputs into manually post-edited counterparts. They use SYSTRAN as the RBMT system, and PORTAGE as SMT system. They report a reduction in post-editing effort of up to a third when compared to the output of the RBMT. Isabelle et al. [7] conclude that an RBMT+SPE system effectively improves the output of a vanilla RBMT system as an alternative to manual adaptations. Experiments show that a SPE system using a corpus with 100,000 words of post-edited translations can outperform a lexicon-enriched baseline RBMT system while reducing the cost.

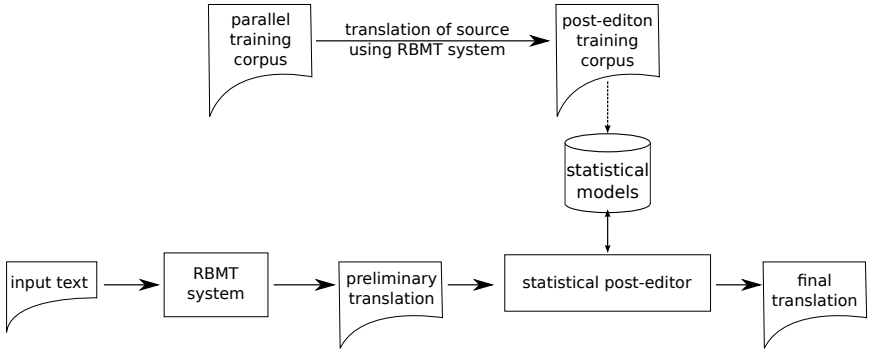

Fig. 2 Architecture of a typical statistical post-editor

Dugast et al. [5] show that a combination of SYSTRAN and an SMT system trained for SPE significantly improves the lexical choice of the final output, even if little improvement is observed in word ordering and grammar. Their comparative

[4] $http://www.commonsenseadvisory.com/Default.aspx?Contenttype=ArticleDetAD\&tabID=63\&Aid=1180\&moduleId=390$

[5] http://lingotek.com

[6] http://www.translingual-europe.eu/slides/WillemStoeller.pdf

[7] http://www.omegat.org

analysis suggests ways to further improve these results by adding "linguistic control" mechanisms. Lagarda et al. [8] show that an SPE system built with the Europarl corpus complements and improves their RBMT system in terms of suitability in a real translation scenario (average improvement of 59.5%). Improvements were less significant (6.5%) for a more complex corpus.

Potet et al. [10] experiment with a small corpus of 175 post-edited sentence pairs (around 5,000 words). These data were used at three different stages of the translation process: (a) extending the training corpus, (b) automatically post-editing the RBMT outputs, and (c) adjusting the weights of the log-linear model. Their experiments show that the use of this small corpus is helpful for correcting and improving the system to retranslate the same data, but it is challenging to propagate these corrections to new data.

Previous experiments for Basque [4] differ from similar work in that a morphological component is used in both RBMT and SMT translations, and in that the size of available corpora is small. The post-edition corpus was artificially created from a bilingual corpora, creating new RBMT translations for the source sentences and taking the corresponding target sentences as *the post-edited sentences* (see Figure 2). They reported improvements when using an RBMT+SPE approach on a restricted domain but a smaller improvement when using more general corpora. In order to improve the MT system, the training material for the post-editing layer of our system consists of a text corpus in two parallel versions: raw machine translation outputs and manually post-edited versions of these translations. Since few resources are available [11], we built the training material from collaboratively constructed language resources.

3 Methodology

This section describe the collaborative post-editing framework. Figure 3 shows the overall architecture of our translation system (RBMT+SPE) that first uses the rule-based system and then the statistical post-edition. Firstly, we describe the aim of the overall system. Secondly, we describe OmegaT, the general human post-editing environment used in this work, as well as the extensions implemented to adapt the tool to the translation of Wikipedia entries. Thirdly, we tweak the translation system to customize it for the Computer Science domain. Finally, the most relevant aspects of the design of the collaboration initiative are described.

3.1 The Aim

The main objective of this work is to build and test an MT system based on the RBMT+SPE approach using manually post-edited corpora from Basque Wikipedia editors. We chose articles in the Computer Science domain, both because it is

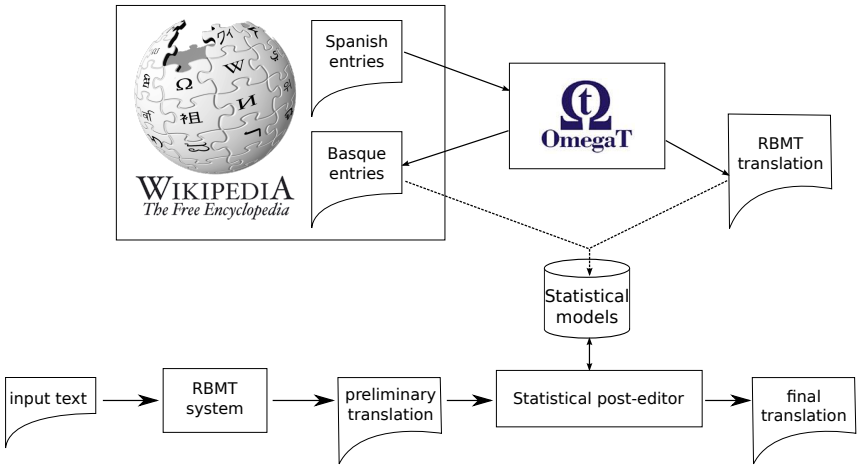

Fig. 3 Architecture of our post-edition environment

suitable as a domain that does not highly depend on cultural factors and because it allows to focus improvements on a domain-specific scenario as a first step.

We expected editors to extend Basque Wikipedia by post-editing Basque RBMT translations of Spanish Wikipedia articles from the Computer Science domain. At the same time, Basque Wikipedia editors would be providing post-edition logs to feed an MT engine. Since Basque and Spanish belong to different language families, we hypothesized that amateur translators (unlike perhaps professional ones) would find the MT output of substantial help.

With the aim of facilitating the post-edition task for editors, we adapted the well-known open-source tool OmegaT. We stored the post-edited translations they provided as a resource to train a SPE system and evaluate the RBMT+SPE engine.

3.2 Modifications to OmegaT

We considered several alternatives to OmegaT when selecting the translation platform to be used during the project, with priority toward open source solutions. We explored a number of tools such as Lokalize, Pootle, Virtaal and OmegaT. Lokalize and Pootle are localization tools that are overly complex for the translation of general texts. Virtaal[8] was initially developed as a specialized tool for translating software but has since moved towards a more graphic-based translation tool. OmegaT is a popular tool among translators and we found interesting features in it that made it suitable for translating general texts. Therefore, OmegaT was selected as the trans-

[8] http://translate.sourceforge.net/wiki/virtaal

lation platform to be used in our project. Other alternatives that were discarded include:

(a) World Wide Lexicon (WWL) Translator, a Firefox add-on that makes browsing foreign-language sites easy and automatic. When browsing a URL, it detects the source language and translates the texts using human and machine translations. Even though it is very useful to navigate through web pages in one's own language, its post-editing interface was not yet fully functional.
(b) Google Translation Toolkit, which provides specific help to translate Wikipedia contents. We had limited access to it as it is not free and open-source tool.

OmegaT is a many-faceted translation application written in Java that, among other advantages, assists translators in their work with translation memories. When a translator imports a text onto OmegaT, the text is segmented for faster and easier reading, while the context of each segment is preserved when a context can aid in translation. From the several features offered by the program, the most useful is the fuzzy matching of text segments against translation memory entries. These matches are displayed when working on a particular text segment, and therefore, the user can easily reuse existing matches for the current segment.

OmegaT also allows to access machine translation systems, in a very similar way to the use of translation memories. The user can choose among several machine translation services (e.g., Google Translate, Apertium and Belazar). The translations produced by the selected systems are shown to the user as alternatives to choose from.

Other features of OmegaT include creating glossaries and dictionaries, importing dictionaries and translation memories, good compatibility with a variety of third-party software, and support for several file types and encodings as well as for different languages. OmegaT is open source and freely available (from `www.omegat.org`), and is supported by extensive documentation and an active community of users and developers.

To make it easier to use for editors, we adapted the interface of OmegaT with a number of additional features:

- *Integration of our Spanish to Basque MT engine.* OmegaT includes a class that connects several machine translation services, making it relatively easy to customize by adding more services (see Figure 4). We used this class to integrate Matxin [9] within OmegaT. In order to reduce the integration effort, we made Matxin's code simpler, lighter and more readable so that it could be implemented as a web service to be accessed by single API calls using SOAP. Therefore, OmegaT could easily make use of a Spanish to Basque machine translation system.
- *Import/export of Wikipedia articles to/from OmegaT.* We implemented a new feature to upload the translated article to the Basque Wikipedia to OmegaT's existing capability of importing MediaWiki documents from their URL encoded as UTF8. To enable this new feature, we also implemented a new login module and some more details. When uploading an article to Wikipedia, the editor is also

required to provide a copy of the translation memory created with the article. We use these translation memories in the process of improving the machine translation service, Matxin. The new upload is language-independent, and can be used for languages other than Basque. However, this feature has not been tested yet on languages that rely on different character sets such as CJK or Arabic.

- *Integration of the Basque spell-checker to facilitate post-editing.* Thanks to OmegaT's flexible support for third-party applications, we also integrated a Basque spell-checker to assist users during translation.
- *Other improvements related to the translation of metadata in Wikipedia.* As an example of translation of Wikipedia metadata, let us take the translation of the internal Wikipedia link [[*gravedad* | *gravedad*]] in the Spanish Wikipedia (equivalent to the link [[*gravity* | *gravity*]] in the English Wikipedia). Our system translates it as [[GRABITAZIO | LARRITASUNA]], so it translates the same word in a different way when it represents the entry Wikipedia and when it is the text shown in such a link. On the one hand, the link to the entry *gravedad* in the Spanish Wikipedia is translated as GRABITAZIO (gravitation) making use of the mechanics of MediaWiki documents which include information on the languages in which a particular entry is available, and their corresponding entries. And on the other hand, the text word *gravedad* is translated as LARRITASUNA (seriousness) using the RBMT system. Therefore, this method provides a translation adapted to Wikipedia. Offering this option allows the post-editor to correct the RBMT translation with the usually more suitable "Wikipedia translation".

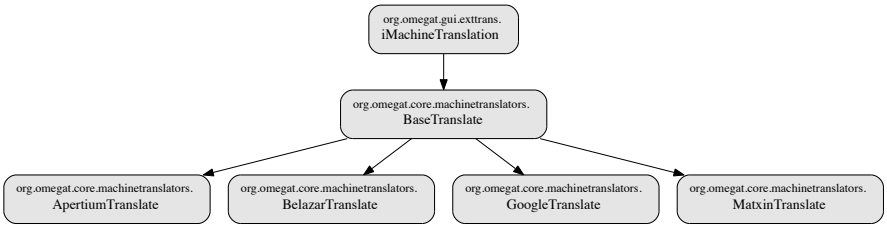

Fig. 4 OmegaT extended with a module to enable the use of the Matxin MT system

The fact that OmegaT needs to be locally installed and configured is inconvenient when the application is going to be used by a large community of users. Our project would have benefited from having access to an on-line contributive platform like Google Translation Toolkit or platforms based on the concept of Interactive Multilingual Access Gateway [3]. To address this shortcoming in existing tools, we are planning to adapt or develop a suitable platform to be used in future projects. Another issue with OmegaT is the somewhat steep learning curve. A new user may feel overwhelmed with the large number of features of the application, and even after gaining a basic familiarity may find it challenging to locate the most appropriate functionalities for the task at hand. Fortunately, there are several tutorials available that help with this. We have also written some user guides to satisfy the needs of

our collaborators; this documentation serves both for understanding the features we have added and for getting the most of the features we deem particularly appropriate for this specific project. In our experience, with some guidance, users quickly overcome initial difficulties, and acquire enough proficiency to work with OmegaT independently.

3.3 Modifications to Matxin RBMT system

The Matxin RBMT system was adapted to the Computer Science domain. The bilingual dictionary was customized in two ways:

Adaptation of lexical resources from dictionary-systems. Using several Spanish/Basque on-line dictionaries, we performed a systematic search for word meanings in the Computer Science domain. We included 1,623 new entries in the lexicon of the original RBMT system. The new terms were mostly multi-words, such as *base de datos* (database) and *lenguaje de programación* (programming language). Some new single words were also obtained; for example, *iterativo* (iterative), *ejecutable* (executable) or *ensamblador* (assembly). In addition, the lexical selection was changed for 184 words: e.g. *rutina*-ERRUTINA (routine) before *rutina*-OHITURA (habit).

Adaptation of the lexicon from a parallel corpus. We collected a parallel corpus in the Computer Science domain from the localized versions of free software from Mozilla, including Firefox and Thunderbird (138,000 segments, 600,000 words in Spanish and 440,000 in Basque). We collected the English/Basque and the English/Spanish localization versions and then generated a new parallel corpus for the Spanish/Basque language pair, now publicly available. These texts may not be suitable for SMT but they are useful for extracting lexical relations. Based on Giza++ alignments, we extracted the list of possible translations as well as the probability of each particular translation for each entry in the corpus. In favour of precision, we limited the use of these lists to the lexical selection. The order was modified in 444 dictionary entries. For example, for the Spanish term *dirección*, the translated word HELBIDE (address) was selected instead of NORABIDE (direction).

3.4 Design of the Collaborative Work

The collaboration between Basque Wikipedia editors and the University of the Basque Country started in 2010. In November 2010 we launched a Wikiproject[9] to collect and disseminate information about the project. Besides the links for downloading the adapted version of OmegaT, the Wikiproject included a list of target arti-

[9] http://eu.wikipedia.org/wiki/Wikiproiektu:OpenMT2_eta_Euskal_Wikipedia

cles to be translated from the Spanish Wikipedia (which had no equivalents available at the time in Basque). These articles were classified by length into three subsets: short (less than 600 words), intermediate (between 600 and 1,000 words) and long (over 1,000 words). The short articles were intended to help editors learn the overall process of downloading an article from Wikipedia, translating and post-editing it, to finally upload the result back to Wikipedia.

Our initial plan was that each editor who "practiced" with a short article would also translate one of the 60 long articles. However, translating a long article represented a substantial amount of work (we estimate that the editors spent more than eight hours translating some long articles). The translation of the 60 long articles was thus taking too long, and therefore we created a tool to help us search for short untranslated Wikipedia entries. This tool is a perl script named *wikigaiak4koa.pl* (*http://www.unibertsitatea.net/blogak/testuak-lantzen/2011/11/22/wikigaiak4koa*) that, given a Wikipedia category and four languages, returns the list of articles contained in the category with their corresponding equivalents in those four languages and their length (1 Kb $\sim$ 1000 characters).

For instance, the following command:

```
$ perl wikigaiak4koa.pl "ca" "eu" "en" "es" "Informática"
```

searches for entries in the "*Informática*" (computer science) category on the Catalan Wikipedia ("ca"), looks for corresponding articles in Basque ("eu"), English ("en") and Spanish ("es"), and finally produces a text file like the following:

```
...

@    31.30 Kb
    eu    A_bildu                      25.25 Kb
    en    At_sign                     113.23 Kb
    es    Arroba_(símbolo)             45.20 Kb
Acord_de_Nivell_de_Servei    22.21 Kb
    en    Service-level_agreement      18.96 Kb
    es    Acuerdo_de_nivel_de_servicio 23.39 Kb
Actic    23.25 Kb
Govern_electrónic    23.69 Kb
    en    E-Government                 18.82 Kb
    es    Gobierno_electrónico         23.18 Kb

...
```

This example examines the Catalan entries for "@," "*Acord_de_Nivell_de_Servei*", "*Actic*" or "*Govern_electrónic*". We can observe that there are equivalent entries in Basque ("*A_bildu*", 25.25 Kb), English ("*At_sign*", 113.23 Kb) and Spanish ("*Arroba_(símbolo)*", 45.20 Kb) and that there is no Basque equivalent for the other three articles in Catalan. The script also shows that these entries are not very long, except the entry for "*At_sign*" in English, which size is 113.23 Kb.

Using this perl script we identified 140 entries that: (1) were included in the Catalan and Spanish Wikipedias, (2) were not in the Basque Wikipedia, and (3) the

size in the Spanish Wikipedia was smaller than 30 Kb ($\sim$ 30,000 characters). These 140 intermediate size entries were included in the Wikiproject. The script can be used to examine the contents of any Wikipedia category for any language.

The size of the Catalan Wikipedia (378,408 articles) is midway between the Spanish (902,113 articles) and the Basque (135,273 articles). Therefore, we consider that a Wikipedia article that is present in the Catalan Wikipedia but not in the Basque Wikipedia should be included in the latter before other non-existing articles that are not in the Catalan version.

4 Results and Discussion

During the first months of 2012 the post-edited texts were processed in order to train a SPE engine and the RBMT+SPE pipeline system was evaluated.

Drawing on previous experience [4] and taking into account the morphology of Basque, we implemented a new automated statistical post-editing system. In this new experiment, the SPE corpus is a real post-edition corpus built from the raw RBMT translation outputs and their corresponding post-editions.

4.1 Evaluation

Tables 1 and 2 show the scoring for different metrics for MT evaluation [12]. The MBLEU, BLEU, NIST and METEOR metrics measure the intersection between the output of the MT system and the human translation; TER, WER and PER express the number of changes necessary to get from the output of the MT system to the human translation. For example, TER (Translation Edit Rate) measures the amount of post-editing that a human would have to perform to change a system output so it exactly matches a reference translation. Possible edits include insertions, deletions, and substitutions of single words as well as shifts of word sequences. All edits have equal cost.

For the former metrics, a higher value represents a higher correlation with human judgments, whereas for the latter metrics lower values are optimal.

The original RBMT system and the RBMT system adapted to the Computer Science domain were tested with the whole set of sentences in selected Spanish Wikipedia articles, and their corresponding sentences after manual correction of RBMT outputs (see Table 1). The improvement is marked for all the metrics when the RBMT system is adapted to the domain. The highest relative improvement is for the BLEU and MBLEU metrics (15%), and the lowest for WER (3.5%).

The final aim of our experiments was to improve the output of the customized RBMT system using statistical post-editing (see Figure 3). However, the corpus used to train the SPE system contained only 50,000 words. This is not an optimal size for statistical training and therefore we had to explore different ways to use the corpus

Table 1 Evaluation of the RBMT systems

SYSTEM	MBLEU	BLEU	NIST	METEOR	TER	WER	PER
Original RBMT	18.89	19.50	6.17	43.94	65.11	68.69	52.08
Adapted RBMT	**21.84**	**22.38**	**6.58**	**47.20**	**62.40**	**66.31**	**49.24**

successfully. We performed a 5-fold cross-validation to evaluate different versions of the RBMT+SPE pipeline optimized by training with different subsets of the post-edition corpus. The sentence pairs were reordered using their TER scores, so that the most similar sentence pairs were promoted to the beginning of the corpus. Three different RBMT+SPE systems were trained using subsets of the corpus with the top 50%, 75% and 100% of this list ordered by TER.

The evaluation on these three systems was repeated for the other three RBMT+SPE systems where the SPE systems were optimized via MERT using a fifth of the corpus (see Table 2).

Table 2 Evaluation of the RBMT+SPE systems

SYSTEM				MBLEU	BLEU	NIST	METEOR	TER	WER	PER
Test	Optim.	Train.								
1/5	1/5	(3/5)	50%	21.57	22.24	6.41	46.48	63.25	67.04	50.36
1/5	1/5	(3/5)	75%	22.54	23.26	6.52	47.55	62.28	66.20	49.59
1/5	1/5	(3/5)	100%	23.66	24.61	6.62	48.06	61.44	65.48	48.98
1/5	0	(4/5)	50%	22.14	22.82	6.50	47.26	62.49	66.56	49.77
1/5	0	(4/5)	75%	23.37	24.10	6.60	48.35	61.67	65.76	49.05
1/5	0	(4/5)	100%	24.24	25.10	6.69	48.94	60.97	65.08	48.58

All the RBMT+SPE systems significantly obtained a better quality than the original RBMT system, and all but the system optimized and trained using only 50% of the corpus achieved a better quality than the customized RBMT system.

The use of a smaller subset of the post-editing corpus with only the most similar sentence pairs produces no improvement in performance; in contrast, a greater number of sentence pairs always leads to improved results, even when the sentence pairs contain greater divergence. This is probably the result of the limited size of our training corpus, and indicates that a larger post-edition corpus might lead to better results.

The best system does not use any subset of the corpus for MERT optimization and uses 100% of the sentences for training. It gets an improvement of 1.82 points for BLEU and 3.4 for MBLEU with respect to the customized RBMT. If compared to the original RBMT system, there is an improvement of 5.6 BLEU points or 5.35 MBLEU points. The other metrics confirm these improvements.

The use of a subset of the corpus for MERT optimization is not a good investment. When using only 50% of the sentences the results are slightly worse, while

using 75% of the sentences only brings a small improvement. Finally, using all the post-edited sentences does produce an improvement, although note that the improvement is higher for the non-optimized system.

4.2 Resources Obtained from the Collaborative Work

When the public collaboration campaign had been running for nine months, from July 2011 to February 2012, 100 new entries and 50,204 words had been added to the Basque Wikipedia. Figure 5 shows the evolution of the number of words translated by editors in that period.

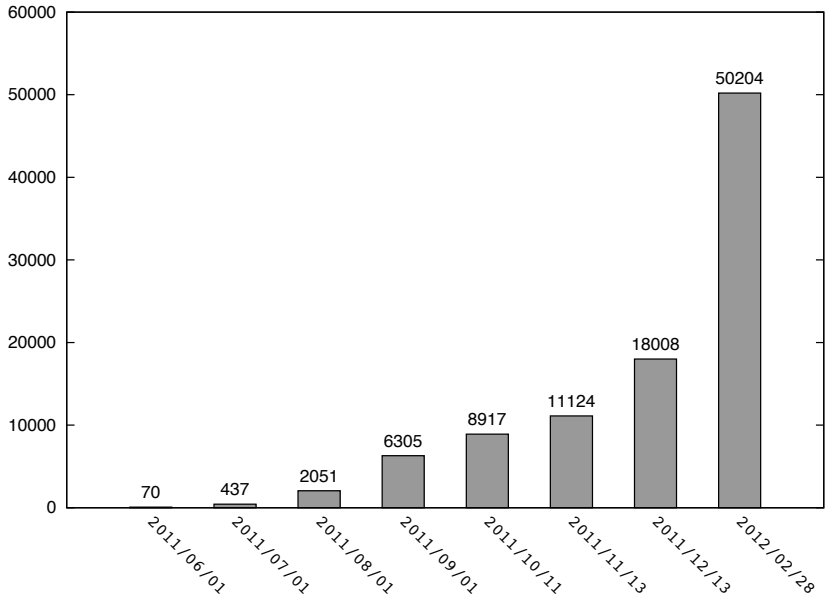

Fig. 5 Evolution of the number of words translated by editors

The current state of our work is described on the web site of the Wikiproject.[10] 100 new entries were added to Basque Wikipedia (the complete list[11] is available looking for articles in Basque Wikipedia defined with the "OpenMT-2" template). and the corpus created by manual post-editing of the RBMT outputs of these new 100 entries is publicly available.[12]

[10] http://eu.wikipedia.org/wiki/Wikiproiektu:OpenMT2_eta_Euskal_Wikipedia

[11] $http://eu.wikipedia.org/w/index.php?title=Berezi:ZerkLotzenDuHona/Txantiloi:OpenMT-2$

[12] http://ixa2.si.ehu.es/glabaka/OmegaT/OpenMT-OmegaT-CS-TM.zip

This data and the interviews with the Wikipedia editors collaborating in the project allow us to draw the following conclusions :

- The use of a MT system, even when its quality is not high, does help editors.
- Short Wikipedia articles are more appropriate to incorporate new collaborators that are sometimes not very motivated to participate in work excessively long.
- More than thirty different users have collaborated in the project so far and almost twenty of them have finished a long article.
- The metadata included in the Wikipedia articles is a challenge for the MT engine and for the users.
- Creating and coordinating a community to produce this type of material in a less resourced language is not an easy task, it can be a substantial task.

4.3 Discussion

An analysis of the post-edited texts helps to better understand the quantitative evaluation, as well as identify the cases where the machine translation works well and can be improved. To perform such an analysis, we first sorted all the translation hypotheses created by the RBMT system (HYP) and their corresponding post-edition outputs (PDT) depending on their TER score. Next, we manually analyzed sentence pairs to validate the usefulness of this corpus. We observed that many of the post-editions with a TER score between 0 and 50 suggested reasonable lexical translation alternatives to the output of the RBMT system. Even though some of those suggestions were for single words, most of them were for multi-word terms. In many cases the post-edited terms appeared with their most frequent inflection suffixes, and that produced several errors. We identified three main problems that can be improved using a statistical post-editing system:

Lexical gaps. When a word is not an entry in the RBMT system's bilingual lexicon this word is not translated by the RBMT system. For example, *video* is not in the lexicon, but its equivalent in Basque (BIDEO) was proposed by editors:

HYP: 3GP **VIDEOA** GORDETZEN DU MPEG - LAU EDO 263 H. . . .
PDT: 3GP **BIDEOA** GORDETZEN DU MPEG - LAU EDO 263 H. . . .

Lexical selection. A better lexical selection can be achieved as a result of training the SPE system with simple contexts. For example, HEDAPEN and LUZAPEN are correct Basque translations for *extensión* (in Spanish). But only LUZAPEN is used for the specific meaning "file extension".

HYP: **HEDAPENA** TXTA HERRITARRA EGIN DA AZKEN GARAIETAN . . .
PDT: TXT **LUZAPENA** ASKO ZABALDU EGIN DA AZKEN GARAIETAN . . .

Ambiguous syntactic structures. Syntactic ambiguities are not always correctly resolved by the parser in the RBMT system. Those corrections that were often registered by the post-editors could be used by the statistical post-editing system to recover correct translations of short chunks. For instance, the translation of *lenguaje*

de restricciones (constraint language) could be translated into Basque as MURRIZTEEN HIZKUNTZA (the language of the constraints) or MURRIZTE HIZKUNTZA (constraint language). Of these, the latter is the correct translation, but the RBMT system provides only the former.

HYP: OCL (object constraint language - OBJEKTUEN **MURRIZTEEN** HIZKUNTZA)

PDT: OCL (object constraint language - OBJEKTUEN **MURRIZTE** HIZKUNTZA)

Depending on the amount of post-editing data, some of these features will be learned by the SPE without the need of modifying the quite complex structure of the RBMT engine. For instance, the first two examples above were properly corrected by the SPE system, while the third one remained unchanged.

5 Conclusions and Future Work

Creating and coordinating a community to produce materials for a less resourced language can be a substantial task. We have defined a collaboration framework that enables Wikipedia editors to generate new articles while they help development of machine translation systems by providing post-edition logs. This collaboration framework has been experimented with editors of Basque Wikipedia. Their post-editing on Computer Science articles were used to train a Spanish to Basque MT system called Matxin. The benefits were twofold: improvement of the outputs of the MT system, and extension the Basque Wikipedia with new articles.

Various auxiliary tools developed as part of this research can also be considered as valuable resources for other collaborative projects: i) a perl script that, given a Wikipedia category and four languages, returns the list of articles contained in the category with equivalents in those four languages and their length. The script is therefore useful to search short untranslated Wikipedia entries; ii) the method used to translate Wikipedia links making use of the mechanics of MediaWiki documents which include information on the languages in which a particular entry is available, and their corresponding entries. This allows the post-editor to correct the RBMT translation with a more suitable "Wikipedia translation"; iii) a new feature added to OmegaT to import/export Wikipedia articles to/from OmegaT. This new upload feature, although used for Basque, was developed as a language-independent tool.

The complete set of publicly available resources created in this project includes the following products:

- The 100 new entries added to Basque Wikipedia.[13]
- The new Spanish/Basque version of the parallel corpus[14] created from the localized versions of free software from Mozilla.

[13] $http://eu.wikipedia.org/w/index.php?title=Berezi:ZerkLotzenDuHona/Txantiloi:OpenMT-2$

[14] http://ixa2.si.ehu.es/glabaka/lokalizazioa.tmx

- The corpus[15] created by manual post-editing of the RBMT outputs of the new 100 entries.
- The perl script *wikigaiak4koa.pl*[16] that returns the list of articles contained in a Wikipedia category (for four languages).
- The improved version of OmegaT,[17] and its user guide.[18]
- The new version of the Matxin RBMT system[19] customized for the domain of Computer Science available as a SOAP service.

We logged the post-editions performed by Wikipedia editors by translating 100 articles from the Spanish Wikipedia into Basque using our MT engine. At the beginning of this work, we set forth the hypothesis that MT could be helpful to amateur translators even if not so much to professionals. After a qualitative evaluation, we can confirm our hypothesis, as even when the quality of the MT output was not high, it was enough to prove useful in helping the editors perform their work. We also observed that Wikipedia metadata makes more complicated both the MT and the post-editing processes, even if the use of Wikipedia's interlanguage links effectively help translation.

Integrating the outcome of collaborative work previously performed in software localization produced a significant enhancement in the adaptation of the RBMT system to the domain of Computer Science. In turn, incorporating the post-editing work of our Wikipedia collaborators into an RBMT system (50,000 words) produced an additional important improvement, despite the fact that the size of this corpus is smaller than those referenced in the major contributions to SPE (for example, Simard et al. [11] used a corpus of 100,000 words). Thus, there may be room for further improvement by the simple expedient of using a larger post-edition corpus. As short Wikipedia articles are more appropriate to incorporate new collaborators, search tools to look for candidate articles in Wikipedia become extremely useful.

The quantitative results show that the contributions can improve the accuracy of a combination of RBMT-SPE pipeline at around 10%, after the post-edition of 50,000 words in the Computer Science domain. We believe that these conclusions can be extended to MT engines involving other less-resourced languages lacking big parallel corpora or frequently updated lexical knowledge.

In addition, the post-editing logs can function in an intermediate fashion as a valuable resource for diagnosing and correcting errors in MT systems, particularly lexical errors that depend on a local context.

Further improvements could be achieved using several tuning techniques. In the near future we plan to study the use of a combination of real post-edition and parallel texts as a learning corpus for SPE.

---

[15] http://ixa2.si.ehu.es/glabaka/OmegaT/OpenMT-OmegaT-CS-TM.zip

[16] http://www.unibertsitatea.net/blogak/testuak-lantzen/2011/11/22/wikigaiak4koa

[17] $http://ixa2.si.ehu.es/glabaka/OmegaT/OpenMT-OmegaT.zip$

[18] $http://siuc01.si.ehu.es/jipsagak/OpenMT\_Wiki/Eskuliburua\_Euwikipedia+Omegat+Matxin.pdf$

[19] $http://ixa2.si.ehu.es/matxin_zerb/translate.cgi$

Acknowledgements This research was supported in part by the Spanish Ministry of Education and Science (OpenMT2, TIN2009-14675-C03-01) and by the Basque Government (Berbatek project, IE09–262). We are indebted to all the collaborators in the project and especially to the editors of the Basque Wikipedia. Elhuyar and Julen Ruiz helped us to collect resources for the customization of the RBMT engine to the domain of Computer Science.

References

[1] Alegria I, Diaz de Ilarraza A, Labaka G, Lersundi M, Mayor A, and Sarasola K (2007) Transfer-based MT from Spanish into Basque: Reusability, Standardization and Open Source. In: *CICLing 2007. LNCS* 4394:374–384. Springer

[2] Alegria I, Diaz de Ilarraza A, Labaka G, Lersundi M, Mayor A, and Sarasola K (2011) Matxin-Informatika: Versión del traductor Matxin adaptada al dominio de la informática. In: *Proceedings of the XXVII Congreso SEPLN*, pp. 321–322. Huelva, Spain

[3] Boitet C, Huynh CP, Nguyen HT, and Bellynck V (2010) The iMAG concept: Multilingual access gateway to an elected web sites with incremental quality increase through collaborative post-edition of MT pretranslations. In: *Proceedings of Traitement Automatique du Langage Naturel* TALN, Montréal

[4] Diaz de Ilarraza A, Labaka G, and Sarasola K (2008) Statistical post-editing: A valuable method in domain adaptation2009) of RBMT Systems. In: *Proceedings of MATMT2008 workshop: Mixing Approaches to Machine Translation*, pp. 35–40. Euskal Herriko Unibersitatea, Donostia

[5] Dugast L, Senellart J, and Koehn P (2007) Statistical post-editing on SYSTRAN's rule-based translation system. In: *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 220–223. Prague

[6] Dugast L, Senellart J, and Koehn P (2009) Statistical post editing and dictionary extraction: Systran/Edinburgh submissions for ACL-WMT2009. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 110–114. Athens

[7] Isabelle P, Goutte C, and Simard M (2007) Domain adaptation of MT systems through automatic post-editing. In: *Proceedings of the MT Summit XI*, pp. 255–261. Copenhagen

[8] Lagarda AL, Alabau V, Casacuberta F, Silva R, and Díaz-de-Liaño E (2009) Statistical post-editing of a rule-based machine translation system. In: *Proceedings of NAACL HLT 2009. Human Language Technologies: The 2009 annual conference of the North American Chapter of the ACL*, Short Papers, pp. 217–220. Boulder

[9] Mayor A, Diaz de Ilarraza A, Labaka G, Lersundi M, and Sarasola K (2011) Matxin, an open-source rule-based machine translation system for Basque. Machine Translation Journal **25**, 1:53–82

[10] Potet M, Esperança-Rodier E, Blanchon H, and Besacier L (2011) Preliminary experiments on using users' post-editions to enhance a SMT system. In: M. L. Forcada, H. Depraetere, V. Vandeghinste (eds.) *Proceedings of the 15th*

Conference of the European Association for Machine Translation, pp. 161–168. Leuven, Belgium

[11] Simard M, Ueffing N, Isabelle P, and Kuhn R (2007) Rule-based translation with statistical phrase-based post-editing. In: *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 203–206. Prague

[12] Snover M, Dorr B, Schwartz R, Micciulla L, and Makhoul J (2007) A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 223–231 Cambridge, Massachusetts, USA

[13] Way A (2010) Machine translation. In: A. Clark, C. Fox, and S. Lappin (eds.) *The Handbook of Computational Linguistics and Natural Language Processing*, pp. 531–573. Oxford: Wiley-Blackwell