# Wikipedia and Machine Translation: killing two birds with one stone

**Iñaki Alegria (1), Unai Cabezon (1) , Unai Fernandez de Betoño (2), Gorka Labaka (1),**

**Aingeru Mayor (1), Kepa Sarasola (1) and Arkaitz Zubiaga (3)**

(1) Ixa Group, University of the Basque Country UPV/EHU,
(2)Basque Wikipedia and University of the Basque Country,
(3)Basque Wikipedia and Applied Intelligence Research Centre of the Dublin Institute of Technology

Informatika Fakultatea, Manuel de Lardizabal 1, 20013 Donostia (Basque Country)

E-mail: i.alegria@ehu.es

## Abstract

In this paper we present the free/open-source language resources for machine translation created in OpenMT-2 wikiproject, a collaboration framework that was tested with editors of Basque Wikipedia. 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 a Rule Based Machine Translation system in nearly 10% benefiting from the post-edition of 50,000 words in the Computer 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.

**Keywords:** collaborative work, Machine Translation, Wikipedia, Statistical Post-Edition

## 1. Introduction

A way for improving Rule Based Machine Translation (RBMT) systems is to use a Statistical Post-Editor (SPE) that automatically post-edits the output of the MT engines. But building a SPE requires a corpus of MT outputs and their manual post-editions pairs.

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.

One of the key features on the success of Wikipedia, the popular and open online encyclopaedia, 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 (Way, 2010). This presents the issue that current machine translation systems generate inaccurate translations that require substantial post-editing by human editors.

In this paper, we introduce our methodology to enable collaboration between Wikipedia editors and researchers, as well as the system we have developed accordingly. This system permits the generation of  new articles by editing machine translation outputs, while editors help improve a machine translation system. We believe that amateur translators can benefit from MT rather than professional translators.

Specifically, to perform such collaboration between editors and researchers, a set of 100 articles were selected from Spanish Wikipedia to be translated into Basque using the machine translation (MT) system called Matxin (Mayor et al., 2011). A group of volunteers from Basque Wikipedia reviewed and corrected these raw translations. In the correction process, they could either post-edit the MT output 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) a set of free/open-source language resources for machine translation, among others the change logs that potentially help improve the MT engine by using an automated statistical post-editor (Simard et al., 2007), and (ii) the generated articles that 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 (Alegria et al., 2013). This improvement was

Section 2 defines the methodology followed in this collaborative project: its design, the criteria and tools used to select the set of Wikipedia articles to be translated, and the resources used to adapt the general MT system to the domain of computer science. Then Section 3 presents the free/open-source language resources and tools created in this project and the achieved translation improvements. The paper ends with the conclusions and future work.

## 2. Related work

Statistical post-editing (SPE) is the process of training a Statistical Machine Translation (SMT) system to translate from rule- based MT (RBMT) outputs into their manually post-edited versions (Simard et al., 2007). They report a reduction in post-editing effort of up to a third when compared to the output of the RBMT system. Isabelle et al. (2007) later confirmed those improvements. A corpus with 100,000 words of post- edited translations outperformed a lexicon-enriched baseline RBMT system.

Using SPE and SYSTRAN as the RBMT system, Dugast et al. (2007, and 2009) significantly improve the lexical choice of the final output. Lagarda et al. (2009) presented an average improvement of 59.5% in a real translation scenario that uses Euoperl corpus, and less significant improvements (6.5 %) when using a more complex corpus.

The first experiments performed for Basque were different because morphological modules were used in both RBMT and SMT translations, and because the size of available corpora was small (Díaz de Ilarraza and al., 2010). The post-edition corpus was artificially created from bilingual corpus: new RBMT translations for the source sentences and taking their target sentences in the bilingual corpus as the post-edited sentences. The improvements they report when using an RBMT+SPE approach on a restricted domain are bigger than when using more general corpora.

Some frameworks for collaborative translation have been created recently. (1) Cross-Lingual Wiki Engine was presented in 2008 (Huberdeau, et al. , 2008). (2) In 2011, the company Asia Online translated 3.5 million articles from English Wikipedia into Thai using MT. (3) Users registered in Yeeyan.org collaboratively translate Wikipedia articles from English to Chinese. (4) Wasala et al. (2013) created a client-server architecture, used in Web localization, to share and use translation memories, which can be used to build (or improve) MT systems. (5) And 'Collaborative Machine Translation for Wikipedia' 1is a Wikimedia proposal for a long-term strategy using several technologies for offering a machine translation system based on collaborative principles. (6) an experiment focused on post-editon of MT output of wiki entries from German and Dutch into English (Gaspari et al., 2011) report that overall the users were satisfied with the system and regarded it as a potentially useful tool to support their work; in particular, they found that the post-editing effort required to attain translated wiki entries in English of publishable quality was lower than translating from scratch.

Popular MT engines include a post-edition interface to fix translations. For instance, Google Translate[2] 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. Other companies such as Lingotek,[3] sell Collaborative Translation Platforms that include post-edition capabilities. For our collaborative work, we use OmegaT, an open source Computer Aided Translation (CAT) tool.

## 3. Design and methodology of the collaborative project on translation

This collaboration among computer scientists, linguists and editors of Basque Wikipedia was developed within the OpenMT-2 Wikiproject. The objective was to design and develop a final MT system by building a Statistical Post-Editor that automatically post-edits the output of the original RBMT system.

To perform such collaboration between editors and researchers, a set of 100 articles from Spanish Wikipedia were translated into Basque using the Matxin RBMT engine. A group of volunteers reviewed and corrected these raw translations. In the correction process, they could either post-edit the MT output to fix errors, or retranslate it when the machine-provided translation was inaccurate. With the aim of facilitating the post-edition task for editors, we adapted the well- known open-source tool OmegaT.

To improve the quality of the Matxin RBMT system's outputs given to the post-editors, we adapted Matxin to the Computer Science domain and the Wikipedia articles to be translated in the project were selected from the Computer Science category. We choose this domain, both because it is suitable as a domain that does not highly depend on cultural factors and because it is a well known domain for our research group.

[1] https://meta.wikimedia.org/wiki/Collaborative_Machine_Translation_for_Wikipedia

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

[3] http://lingotek.com

The public collaboration campaign was run for eight months, from July 2011 to February 2012 and 36 volunteers collaborated in it. This process ultimately produced two main benefits:

1. The raw and manual post-edited translation pairs served to built an automated Statistical Post-Editor. This SPE system can improve the accuracy of the RBMT system in nearly 10%. MBLEU, BLEU, NIST, METEOR, TER,WER and PER metrics confirm this improvement (Alegria et al, 2013).
2. The generated articles help expand the Basque Wikipedia. 100 new entries (50,204 words) had been added to the Basque Wikipedia.

Additionally, improvements have been made in both Matxin and OmegaT systems.

### 3.1 Selection of Wikipedia articles

To incorporate new collaborators that are sometimes not very motivated to participate in work excessively long we decided to translate short Wikipedia articles.

We created a tool to help us search for short untranslated Wikipedia entries. This tool is a perl script named wikigaiak4koa.pl 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.

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.

Using the tool 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 (~ 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.

### 3.2 Modifications to Matxin RBMT system

The Basque-Spanish Matxin RBMT system was adapted to the Computer Science domain. The bilingual lexicon 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

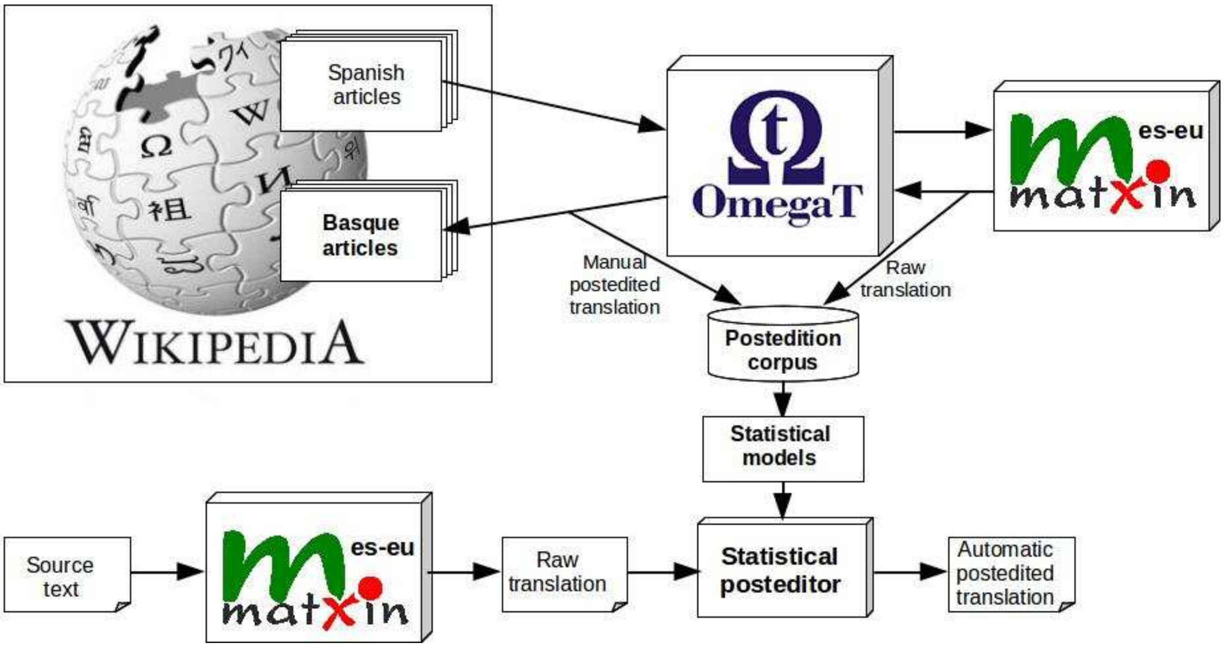

**Figure 1. Architecture of the final MT system enriched with a Statistical Posteditor.**

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).

### 3.3 Modifications to OmegaT

OmegaT was selected as the post-edition platform to be used in our project. To make it easier to use for editors, we adapted the interface of OmegaT with a number of additional features:

- Integration of Matxin 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. We used this class to integrate Matxin 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.
- Integration of the Basque speller, to facilitate post-editing.
- A functionality to 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 building the SPE system.
- A tool for translating Wikipedia links. This module use Wikipedia metadata to search the Basque article that corresponds to a Spanish one. 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".

## 4. Created resources and achieved improvements

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

- Corpus
  - The new Spanish/Basque version of the parallel corpus[4]. created from the localized versions of free software from Mozilla (138.000 segments, 600.000 word in Spanish and 440.000 in Basque).
  - The corpus[5] of raw and manual post-edited translations (50.204 words). It was created by manual post-editing of the Basque outputs given by Matxin RBMT system translating 100 entries from the Spanish Wikipedia.
- Wikipedia
  - The 100 new entries[6] added to Basque Wikipedia (50.204 words).
  - A tool for searching articles in the Wikipedia (wikigaiak4koa.pl[7]). This tool is a perl script that can be used to browse the content of a category for any language in Wikipedia. Given a Wikipedia category and four languages, it returns the list of articles contained in the category with their corresponding equivalents in those four languages and their length.
- Matxin

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

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

[6] http://eu.wikipedia.org/w/index.php?title=Berezi:ZerkLotzenDuHona/Txantiloi:OpenMT-2&limit=250

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

- The new version of the Matxin RBMT system customized for the domain of Computer Science available as a SOAP service.[8]
- A new automated Statistical Post-Editing system. This system has been built using the corpus of raw RBMT translation outputs and their corresponding manual post-editions (50.204 words).
- The quantitative results show that the combination of RBMT-SPE pipeline can improve the accuracy of the raw RBMT system at around 10%, despite the fact that the size of the corpus used to built the SPE system is smaller than those referenced in the major contributions to SPE (for example, Simard et al. 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.

- OmegaT
  - Integration of Matxin Spanish to Basque MT engine.
  - Integration of the Basque speller.
  - A functionality to import/export of Wikipedia articles to/from OmegaT. This 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.
  - A tool for translating Wikipedia links. This module use Wikipedia metadata to search the Basque article that corresponds to a Spanish one.
  - A tutorial in Basque to download, install and use OmegaT, with details to post-edit Wikipedia articles[9].

## 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 SPE system that improves the output of the Spanish to Basque MT system called Matxin.

We set forth the hypothesis that MT could be helpful to amateur translators even if not so much to professionals. 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.

The benefits of this project were twofold: improvement of the outputs of the MT system, and extension the Basque Wikipedia with new articles. Various auxiliary tools and language resources developed as part of this research can also be considered as valuable resources for other collaborative projects.

## 6. References

Alegria I., Cabezon U., Fernandez de Betoño U., Labaka G., Mayor A., Sarasola K., Zubiaga A. (2013) Reciprocal Enrichment between Basque Wikipedia and Machine Translators. In I. Gurevych and J. Kim (Eds.) The People's Web Meets NLP: Collaboratively Constructed Language Resources', Springer. ISBN-10: 3642350844, pp. 101-118.

Diaz de Ilarraza A., Labaka G., Sarasola K. (2008) Statistical post-editing: a valuable method in domain adaptation of RBMT systems. In: Proceedings of MATMT2008 workshop: mixing approaches to machine translation, Euskal Herriko Unibersitatea, Donostia, pp 35–40

Dugast L., Senellart J., Koehn P. (2007) Statistical post-editing on SYSTRAN's rule-based translation system. In: Proceedings of the second workshop on statistical machine translation, Prague, pp 220–223

Dugast L., Senellart J., 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, Athens, pp 110–114

Gaspari F., Toral A., and Naskar S. K. (2011) User-focused Task-oriented MT Evaluation for Wikis: A Case Study.. In *Proceedings of the Third Joint EM+/CNGL Workshop "Bringing MT to the User:* Research *Meets Translators"*. European Commission, Luxembourg, 14 October 2011. 13-22

Huberdeau L.F., Paquet B., Desilets A. (2008). The Cross-Lingual Wiki Engine: enabling collaboration across language barriers. In *Proceedings of the 4th International Symposium on Wikis* (WikiSym '08).

[8] htt p : //ixa2.si.ehu.es/matxin erb/translate.cgi

[9] http://siuc01.si.ehu.es/~jipsagak/OpenMT_Wiki/ Eskuliburua_Euwikipedia+Omegat+Matxin.pdf

ACM, New York, NY, USA

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

Lagarda A.L., Alabau V., Casacuberta F., Silva R., 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, Boulder, pp 217–220

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

Simard M., Ueffing N., Isabelle P., Kuhn R. (2007) Rule-based translation with statistical phrase- based post-editing. In: Proceedings of the second workshop on statistical machine translation, Prague, pp 203–206

Wasala A., Schäler R., Buckley J., Weerasinghe R., Exton C. (2013) Building Multilingual Language Resources in Web Localisation: A Crowdsourcing Approach.. In I. Gurevych and J. Kim (Eds.) The People’s Web Meets NLP: Collaboratively Constructed Language Resources', Springer. ISBN-10: 3642350844, pp. 69-100.

Way A. (2010) Machine translation. In: Clark A, Fox C, Lappin S (eds) The handbook . of computational linguistics and natural language processing. Wiley-Blackwell, Oxford, pp 531–573