Domain Adaptation in MT using Wikipedia as a Parallel Corpus: Resources and Evaluation

Gorka Labaka, Iñaki Alegria, Kepa Sarasola
University of the Basque Country
{gorka.labaka i.alegria kepa.sarasola}@ehu.eus

Abstract

This paper presents how a state-of-the-art Statistical Machine Translation system is enriched by using extra in-domain parallel corpora extracted from Wikipedia. We collect corpora from parallel titles and from parallel fragments in comparable articles from Wikipedia editions for English, Spanish and Basque. We carried out an evaluation with a double objective: to evaluate the quality of the extracted data and to evaluate the improvement from using domain-adaptation. We think this enrichment method can be very useful for languages with limited amount of parallel corpora, where in-domain data is crucial to improve the performance of MT systems. The experiments on the Spanish–English language pair improve a baseline trained on the Europarl corpus in more than 2 BLEU points when translating texts from the Computer Science domain.

Keywords: machine translation, domain adaptation, Wikipedia

1. Introduction

Domain adaptation has recently gained interest within statistical machine translation (SMT) as a way to cope with the performance drop observed when testing conditions deviate from training conditions. The basic idea is that in-domain training data can be exploited to adapt all components of an already developed system. However, previous work showed small performance gains after adaptation using limited in-domain bilingual data (Bertoldi and Federico (2009), Daumé III and Jagarlamudi (2011)).

Our system is based on enriching a state-of-the-art SMT system with in-domain parallel corpora extracted from Wikipedia. We carried out an evaluation with a double objective: to evaluate the quality of the extracted data and to evaluate the improvement from the domain-adaptation method.

We think this kind of domain adaptation can be very useful for languages with limited amount of parallel corpora, where in-domain data is crucial to improve the performance of MT.

As a previous step, comparable corpora were extracted from Wikipedia using a graph-based selection (Barrón-Cedeño et al., 2015). Then the titles of the collected articles were used as parallel sentences to improve a SMT baseline for the Spanish–English and Basque–English pairs.

Finally, for the Spanish–English pair, fragments including parallel sentences are also extracted from the comparable articles in Wikipedia and their contribution to MT evaluated.

The paper is organized as follows. Section 2 introduces several works on this subject and related topics. In Section 3 we summarize from Barrón-Cedeño et al. (2015) the methodology used to extract comparable corpora from Wikipedia. In Section 4 we describe how the parallel subcorpora obtained from them are used to feed the SMT translators, which are then evaluated. Finally, we draw the conclusions and show how this work is being continued.

2. Related work

Some effort has been devoted to improving MT using Wikipedia but generally domain-related parallel corpora from Wikipedia are not identified and only small units, mainly named-entities, are used.

Wikipedia has been used mainly just with the aim of augmenting dictionaries (Jones et al. (2008), Nothman et al. (2013)).

Gupta et al. (2013)) extract parallel fragments of texts from a comparable corpus built from Wikipedia documents, but they do it in general, without restriction to an specific domain. They improve the performance of an existing machine translation system.

Looking at the problem of extracting parallel sentences from Wikipedia Adafre and De Rijke (2006) described two methods to extract parallel (similar) sentences directly.

Patry and Langlais (2011) use a classifier with content-based features (hapax words, numerical entities and punctuation marks) to extract parallel corpora from Wikipedia.

The toolkit Accurat can be applied to Wikipedia to extract a general comparable corpus, operating on different levels (Pinnis et al., 2012). Similarly, CorpusPedia (Otero and López, 2010) is a tool to extract comparable corpora from Wikipedia by considering a pair of languages and a category.

For our aim we decided to use and evaluate the proposal from Barrón-Cedeño et al. (2015), since their method allows us to exploit the category graph information to identify domain specific articles (see next section).

3. Wikipedia as Source for Domain Corpora

Wikipedia's users can categorize articles by including one or more labels in the page's markup. This way, articles are grouped by categories and the category hierarchy forms a graph. However, many articles are not associated to the categories they should belong to. Besides, since users have the freedom to give any category name, the set of categories assigned to an article can be very wide. Plamada and Volk (2012) already demonstrated the difficulty in using

Table 1: Size in number of articles and categories for the three Wikipedia editions used.

Wikipedia edition	Articles	Categories	Ratio
English	4,514,317	1,206,065	3.7
Spanish	1,070,407	261,681	4.1
Basque	249,400	44,879	5.6

Wikipedia categories for the extraction of domain-specific articles from Wikipedia.
Therefore, in order to select articles from a given domain, we have taken advantage from the previous work carried out by Barrón-Cedeño et al. (2015) which is based on the graph structure of Wikipedia to identify domain articles, and uses a wide range of measures for extracting parallel sentences from those articles.
Three steps are taken to obtain a parallel corpus for a domain[1]:

- Selection of monolingual domain-based articles
- Collection of parallel titles via inter-language links
- Extraction of parallel texts inside the articles

In this work we take into account the titles of the extracted articles and categories for the en-es and en-eu pairs and also extract parallel sentences for the en-es pair.
Titles ensure good precision (and low recall) although some noise is possible caused by out-of-domain entries or erroneous inter-language links.

3.1. Domain-based Article Selection

The corresponding Wikipedia dumps for English, Spanish and Basque were downloaded from the Wikimedia Downloads page[2] during January and February 2015 and preprocessed using the JWPL library (Zesch et al. (2008)).
Table 1 summarizes the size of each of the three Wikipedia editions as measured by the number of articles and categories after a first treatment of disambiguation and redirect pages are applied.
Starting from a root category, the model performs a breadth-first search on Wikipedia.
The core idea is to score the explored categories in order to assess how likely it is that they actually belong to the desired area. In the approach, a naïve assumption is made: a category belongs to the area if its title contains at least a word of the domain vocabulary. These domain vocabulary is automatically built from the root category and it is composed by the tokens in its articles, considering only the most frequent 10% of non-stopword tokens.
When exploring the graph of categories, each level is scored by measuring the percentage of categories in it that are associated to the domain by means of this vocabulary.

[1]A more detailed information can be consulted in Barrón-Cedeño et al. (2015)

[2]https://dumps.wikimedia.org

Table 2: Selected depth per category and number of articles at the corresponding depth.

	Computer Science		Science	
	#art.	depth	#art.	depth
English	155,533	7	785,642	6
Spanish	29,634	6	820,949	8
Basque	1,717	6	7,414	4

Two different root categories were selected for evaluation: *Computer Science* and *Science*. Table 2 summarizes the number of articles obtained for each language and domain.

3.2. Parallel Titles: Intersection and Union

The monolingual corpora have been extracted independently for every language, and therefore, an article can be present in one collection but its equivalent article absent in the collection for another language. Two different ways of joining the monolingual corpora have been tested: the intersection and the union of their elements. With the union, the set is larger at the cost of including the noise that the extraction method produces.
Table 3 shows the number of titles extracted for each language-pair, domain and combination method.
In the MT experiments parallel titles from the intersection and the union have been used in order to improve the SMT systems.

3.3. Extraction of parallel sentences

The text in the collections of articles in the three languages form the comparable corpora in the Computer Science and Science domains.
Given a pair of articles related by an inter-language link, the similarity between all their pairs of cross-lingual sentences are estimated using different text similarity measures. The used measures are: (i) character n-grams (cng), (ii) pseudo-cognates and (iii) common words after translation from one language to the other. A length factor (len) is added as a penalty.
After defining a threshold for each measure, the sentence pairs with a similarity higher than the threshold are extracted as parallel sentences. Authors report a lot of experiments and conclude that averaging all the similarities after multiplying them by the length factor is a good option, although it is better to combine the previous with 3grams, pseudo-cognates and translation into English. Following these method two different corpora were extracted from the text in the articles included in the intersection of article titles extracted in Section 3.2. Table 4 shows the figures of the extracted corpora, that we use in our experiments (en-es pair).

4. Evaluation in MT

After the description of the extraction of parallel corpora related to the domain, we describe here the adaptation of a state-of-the-art SMT engine to the Computer Science domain.

Table 3: Number of parallel titles from articles and categories obtained as the union or the intersection of the documents belonging to the same category in the two languages.

	Computer Science			Science		
	#art.	#cat.	#total	#art.	#cat.	#total
en∩es	16,983	1,060	18,043	132,159	9,796	141,955
en∪es	36,843	2,564	39,407	720,715	90,992	811,707
en∩eu	1,025	123	1,148	4,537	517	5,054
en∪eu	4,322	385	4,707	53,919	3,014	56,933

Table 4: Size of the en-es parallel corpora obtained from fragments of the articles.

System		Sentences	tokens
Computer Sc.	en	508,789	10,908,611
	es		11,622,143
Comp. Sc. and Science	en	4,522,339	85,899,900
	es		93,160,232

Table 5: Size of the parallel corpora used in the baseline SMT systems.

System		Sentences	tokens
Europarl	en	1,965,734	55,031,614
	es		57,139,119
Elhuyar	en	1,290,501	16,280,750
	eu		14,952,275

4.1. Experimental Setup: Corpora

The evaluation corpus is the development set used in the QTLeap project (Branco and Osenova, 2014; Agirre et al., 2015). The QTLeap project (Quality Translation by Deep Language Engineering Approaches) is a collaborative project that aims to produce high-quality outbound Machine Translation using deep language engineering approaches to achieve higher quality translations.

The evaluation bench used in that project was provided by *Higher Functions*, and is based on applying Machine Translation to allow multilingual "online consultancy" which handles common problems in computer use. On the multilingual configuration of this application, users ask their questions in their own language. The system translates them into English and retrieves the answer in a data-base of previously answered questions (all of them in English). If a suitable answer is found, the system answers with the result of its translation from English to the user's language (including Spanish and Basque).

The baseline Spanish–English system was trained on the corpora available from the WMT workshop series[3]. Translation models were trained on the Europarl corpus, while for the monolingual language modeling the News Crawl corpus was used (from 2007 to 2012). These baseline systems were extended using the domain corpora in two ways: (1) only using the parallel titles extracted from Wikipedia, and (2) also incorporating the parallel sentences extracted from the articles. Besides we wanted to test the effect of adding information of the science domain, which is bigger but not so closely-related to the test domain.

For the Basque–English baseline system, the corpora used was provided by *Elhuyar*[4]. The corpora consist of several Translation Memories from different clients thus covering a wide rage of topics. It is important to notice that the Computer Science domain is also partially covered, since some of the Translation Memories contain software localization texts. Monolingual corpora for language modeling was extended with additional Basque text made available by the same company. Parallel sentences from Wikipedia for Basque were not available. So, the baseline system was extended using only the parallel titles extracted from Wikipedia. We incorporated the Computer Science domain titles alone or in combination with the Science domain titles.

For evaluation and parameter optimization, we used two of the datasets developed in the QTLeap project. Each dataset (development and test) consists of 1,000 questions and their corresponding answers, therefore 2,000 segments, translated by professional translators from English into Basque and Spanish. Each segment is made up of more than one sentence. Below you can see an illustrative example of an English interaction:

- *In openoffice, how do I insert a chart into a text document?*
- *Click on the part of the document where you want the chart and then in the Insert menu choose Object and click where it says graph.*

4.2. SMT: Training, Tuning and Results

The development of all the systems was carried out using publicly available state-of-the-art tools: the mGIZA toolkit, the SRILM toolkit and the Moses decoder. More concretely, we followed the phrase-based approach with standard parameters: a maximum length of 80 tokens per sentence, translation probabilities in both directions with Good Turing discounting, word-based translation probabilities (lexical model, in both directions), a phrase length penalty and the target language model. The weights were

[3] http://statmt.org/wmt15/

[4] https://www.elhuyar.eus/en

Table 6: Results of MT quality using BLEU for the en-es pairs enriched with titles obtained by the intersection and union of the domain titles of each language.

Pair	System	BLEU
es-en	Europarl (baseline)	23.77
	+WP-inters.(Comp.Sc.)	23.22
	+WP-inters.(Comp.Sc., Science)	23.42
	+WP-union(Comp.Sc.)	**24.04**
	+WP-union(Comp.Sc., Science)	23.36
en-es	Europarl (baseline)	21.80
	+WP-inters(Comp.Sc.)	21.30
	+WP-inters(Comp.Sc., Science)	21.78
	+WP-union(Comp.Sc.)	**22.22**
	+WP-union(Comp.Sc., Science)	21.64

Table 7: Results of MT quality using BLEU for en-es pairs enriched with parallel fragments.

Pair	System	BLEU
es-en	Europarl (baseline)	23.77
	+WP(Computer Sc.)	25.46
	+WP(Computer Sc., Science)	**25.91**
en-es	Europarl (baseline)	21.80
	+WP(Computer Sc.)	**24.46**
	+WP(Computer Sc., Science)	23.89

adjusted using MERT tuning with n-best list of size 100. For the monolingual LM the Europarl and News Crawl corpora (from 2007 to 2012) are interpolated for all the systems.

Table 6 summarizes the results of the evaluation of the en-es and es-en systems where only titles are added. We can see that the incorporation of the titles gets sometimes a small improvement with respect to the baseline system.

Table 7 summarizes the results of the evaluation of the en-es and es-en systems where titles and parallel fragments are added. In this case the improvement is solid.

Table 8 summarizes the results of the evaluation of the English–Basque pairs systems (en-eu and eu-en) where only titles are added, leading to similar conclusions as for the en-es pair (Table 6).

5. Conclusions and Future Work

The experiments on the Spanish–English language pair show that an enriched system improves a baseline trained with the Europarl corpus in more than 2 points of BLEU when all the extracted parallel sentences are used for training (in both cases: using Computer Science domain only or with Science domain). But when only the titles are used only a small improvement is reached. We confirm the poor results when adding only titles in the English–Basque pair. These results can lead to conclusions as follows:

- The tested methods for selecting articles and extracting parallel fragments from Wikipedia are adequate.

Table 8: Results of MT quality using BLEU for English–Basque pairs enriched with titles obtained by the intersection and union of the domain titles of each language.

Pair	System	BLEU
eu-en	Baseline	18.15
	+WP-inters(Comp.Sc.)	18.11
	+WP-inters(Comp.Sc., Science)	17.78
	+WP-union(Comp.Sc.)	18.05
	+WP-union(Comp.Sc., Science)	**18.40**
en-eu	Baseline	13.31
	+WP-inters.(Comp.Sc.)	13.41
	+WP-inters.(Comp.Sc., Science)	**13.57**
	+WP-union(Comp.Sc.)	13.29
	+WP-union(Comp.Sc., Science)	13.26

- Using titles only is not sufficient to get solid improvements in the quality of the baseline SMT systems. Two hypotheses can be contemplated to explain this: (i) the insufficient size of the new title corpus; (ii) the extremely short context present in article titles.

In the near future we aim to investigate better methods for take advantage of Wikipedia as a source to collect resources for domain adaptation in MT.

6. Acknowledgments

We are indebted to Cristina España-Bonet and to Alberto Barrón-Cedeño for their collaboration providing the corpora extracted from Wikipedia.

The research leading to these results was carried out as part of the TACARDI and TADEEP projects (Spanish Ministry of Economy and Competitiveness, TIN2012-38523-C02-011 and TIN2015-70214-P, with FEDER funding) and the QtLeap project (Quality Translation by Deep Language Engineering Approaches, European Comission, Strep, FP7-ICT-2013.4.1-610516).

References

Adafre, S. F. and De Rijke, M. (2006). Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Agirre, E., Alegria, I., Aranberri, N., Artetxe, M., Barrena, A., Branco, A., de Ilarraza, A. D., Gojenola, K., Labaka, G., Otegi, A., et al. (2015). Lexical semantics, basque and spanish in qtleap: Quality translation by deep language engineering approaches. *Procesamiento del Lenguaje Natural*, 55:169–172.

Barrón-Cedeño, A., España Bonet, C., Boldoba, J., and Màrquez, L. (2015). A factory of comparable corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13, Beijing, China. Association for Computational Linguistics.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189. Association for Computational Linguistics.

Branco, A. and Osenova, P. (2014). Qtleap-quality translation with deep language engineering approaches. In *Poster at EAMT2014, Dubrovnik*.

Daumé III, H. and Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 407–412. Association for Computational Linguistics.

Gupta, R., Pal, S., and Bandyopadhyay, S. (2013). Improving mt system using extracted parallel fragments of text from comparable corpora. In *proceedings of 6th workshop of Building and Using Comparable Corpora (BUCC), ACL, Sofia, Bulgaria*, pages 69–76.

Jones, G. J., Fantino, F., Newman, E., and Zhang, Y. (2008). Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from wikipedia. In *proceedings of the 2nd workshop oCross-Lingual Information Access*.

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Otero, P. G. and López, I. G. (2010). Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pages 21–25.

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95. Association for Computational Linguistics.

Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Babych, B., et al. (2012). Accurat toolkit for multi-level alignment and information extraction from comparable corpora. In *Proceedings of the ACL 2012 System Demonstrations*, pages 91–96. Association for Computational Linguistics.

Plamada, M. and Volk, M. (2012). Towards a wikipedia-extracted alpine corpus. In *The 5th Workshop on Building and Using Comparable Corpora*, page 81. Citeseer.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wikictionary. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).