Adapting NMT to caption translation in Wikimedia Commons for low-resource languages

Adaptando NMT a la traducción de pies de imagen en Wikimedia Commons para idiomas con pocos recursos

Alberto Poncelas[1], **Kepa Sarasola**[2], **Meghan Dowling**[1], **Andy Way**[1], **Gorka Labaka**[2], **Iñaki Alegria**[2]
[1]ADAPT Centre, School of Computing, Dublin City University
[2]Ixa Group Faculty of Informatics, (UPV/EHU)
Información de contacto: kepa.sarasola@ehu.eus

Abstract: This paper presents a successful domain adaptation of a general neural machine translation (NMT) system using a bilingual corpus created with captions for images in Wikimedia Commons for the Spanish-Basque and English-Irish pairs.
Keywords: Machine Translation, Low-resource languages, Bilingual corpora, Language resources from Wikipedia

Resumen: Este artículo presenta una adaptación a dominio exitosa de un sistema de Traducción automática neuronal (NMT) utilizando un corpus bilingüe creado con los pies de imagen utilizados en Wikimedia Commons para los pares de idiomas español-euskera e inglés-irlandés.
Palabras clave: Traducción automática, Idiomas con recursos limitados, Corpus bilingüe, Recursos lingüísticos extraídos de Wikipedia

1 Introduction

Wikimedia (the umbrella organisation which includes Wikipedia, Wikidata, Commons, etc.) is one of the most valuable sources for the collection of low-resource language corpora. Due to most of the text-based content in Wikimedia being in majority languages such as English and Spanish, the development of a translation tool could help generate new content in these low-resource languages.

In this paper we try to take advantage of existing text in WikiMedia in order to improve translation quality for low-resource languages. Our hypothesis is that given a corpus created with available bilingual captions for images in Wikimedia Commons (in Spanish and Basque), a successful adaptation of a general NMT translation-system to this domain could improve the quality of the translation of new captions. We also hypothesise that additional semantic information extractable from the Wikipedia context in which the images are being used could also be help improve translation quality.

Using new corpora that we generate from Wikimedia Commons image captions, we demonstrate that significant increments in translation quality can be obtained. Moreover, additional small improvements were detected when using additional semantic tags extracted from Wikipedia or Wikidata.

Following on from this, we hypothesise that this process could be applied not only for Basque, but also to other low-resource languages too. To investigate this, we perform similar experiments for the English-Irish pair.

The aim of these experiments is that the newly improved MT systems created will help to increase the number of image captions for low-resource languages in Commons and over time the increased training corpus size will lead to the creation of an even better MT system.

The paper is organised as follows. Section 2 introduces related work on this subject. In Section 3, we describe in detail the process used to create the parallel corpora. Section 4 details the steps performed in this experiment: creation of the initial corpus, design and evaluation of the baseline and the domain-adapted NMT systems, and then the enrichment of the corpus with semantic tags and the final evaluation. Section 5 shows how this work is being continued with other language pairs, detailing a similar experiment that indicates the extensibility of this work to the English-Irish pair. Finally, in Section 6, we draw some conclusions and propose some lines for future work.

2 Related work

NMT is the dominant paradigm in the field as evidenced in particular by large translation

providers turning to NMT for their production engines (Wu et al., 2016; Crego et al., 2016) and NMT systems achieving the best results in most cases on standard shared task datasets (Bojar et al., 2017).

2.1 Domain adaptation for NMT systems

Fine tuning is the conventional method of domain adaptation in NMT. This consists of using in-domain data for training the last epoch of a pre-built NMT model. Previous work (Luong and Manning, 2015; Poncelas, de Buy Wenniger, and Way, 2018b) has shown that using in-domain data for the last epochs of an NMT system leads to better results.

2.2 Previous results for the es-eu and en-ga pairs

The only reported results for Spanish-Basque (es-eu) using NMT are those described by Etchegoyhen et al. (2018). They report improvements over other methods using NMT and BLEU scores of around 20 and 23 points for two references of the same corpus. In previous work, Labaka et al. (2014) reported their results using statistical machine translation (SMT) and a hybrid model for the same language pair. Other works involving Basque includes the work of Poncelas, Way, and Sarasola (2018) in which Basque-to-English MT models were improved by fine-tuning with in-domain sentences (retrieved using a approximated target side (Poncelas, de Buy Wenniger, and Way, 2018a)).

In terms of the English-Irish (en-ga) pair, Dowling et al. (2015) have previously presented work on domain-specific SMT, and more recently, preliminary experiments involving en-ga NMT (Dowling et al., 2018).

2.3 Wikipedia and MT

In relation to resources for MT, Otero and López (2010) underline the value of Wikipedia as a multilingual source of comparable corpora. Nothman et al. (2013) generate multilingual named entity recognition by exploiting the text and structure of Wikipedia. This cross-lingual approach achieves up to 95% accuracy.

In Labaka, Alegria, and Sarasola (2016) a SMT system is enriched with extra in-domain parallel corpora compiled from parallel titles in Wikipedia. On the Spanish–English language pair they improve a baseline trained on the Europarl corpus by more than 2 BLEU points when translating texts from the Computer Science domain. In our previous work (Alegria et al., 2013) we described a collaboration framework that enables Wikipedia editors to generate new articles while helping the development of MT systems by providing post-editing logs. This collaboration framework was tested with editors of Basque Wikipedia.

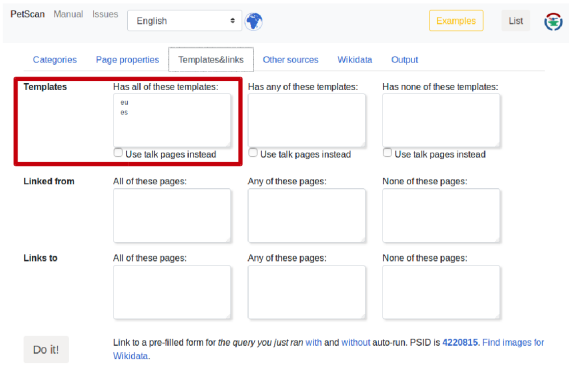

Figure 1: Searching with PetScan for images with templates 'es' (Spanish) and 'eu' (Basque).

3 *Creation of the new corpora*

Our first aim was to build a bilingual corpus, as big as possible, by collecting the captions of images in Wikimedia Commons[1]. We decide to use Petscan[2] for this purpose. The method we have developed can be used for any other pair of the 278 languages present in Wikipedia[3].

3.1 *CommonsCaptions* corpus

Petscan is a useful tool for searching for information in the three main Wikimedia services: Wikipedia (articles), Wikidata (linked data) and Wikimedia Commons (images). Besides a lot of other search capabilities, it also allows the user to look for images in Commons that contain a text description written in a given language, e.g.: searching the {{es}} template provides captions in Spanish. Figures 1 and 2 show two screenshots of the Petscan query that in April 2018 provided us with a list of the 6,876 images with captions in both Spanish and Basque. The list of files with both these templates had some false positives (since the templates defined to introduce text in Basque and Spanish may be used for other fields, since their use is not exclusive to captions).

As a first step our bilingual corpus was created solely through the collection of image cap-

[1] https://commons.wikimedia.org
[2] https://petscan.wmflabs.org
[3] https://stats.wikimedia.org/EN/Sitemap.htm

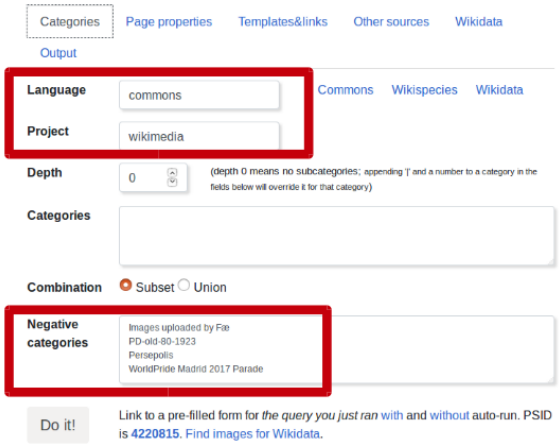

Figure 2: Searching in PetScan for images in Wikimedia Commons (the tab *Categories*).

tions, but eventually extra sentences could be obtained from text used in Commons categories and transclusions (the inclusion of an image into another Wikimedia document by cross-reference).

The caption descriptions detailed here are usually not full sentences. They accompany usages on projects such as Basque and Spanish Wikipedia and, for example, may only contain a single surname if in a biography. It is difficult to collect enough text for a low-resource language like Basque to be able to train an MT system.

However, many of the descriptions generated contained the same general short note which was not an accurate caption. We discarded such images by adding them to a *negative categories* list (see Figure 2). The final query is represented with PSID=4220815 [4], and provided us with 6,876 results.

Finally, after discarding 3,316 images with problematic captions (non-alphabetical captions, non-parallel captions, etc.), the final parallel corpus was composed of a total number of 3,560 captions. This *CommonsCaptions* corpus is publicly available online.[5]

3.2 TaggedCommonsCaptions corpus

After building the *CommonsCaptions* corpus collect semantic information inherent to the Wikipedia context to test whether the inclusion of this knowledge could bring an improvement in the quality of the translations.

We used a simple approach: adding a tag to each of the images to distinguish 11

[4] https://petscan.wmflabs.org/?psid=4220815

[5] http://ixa.si.ehu.eus/node/11513?language=en

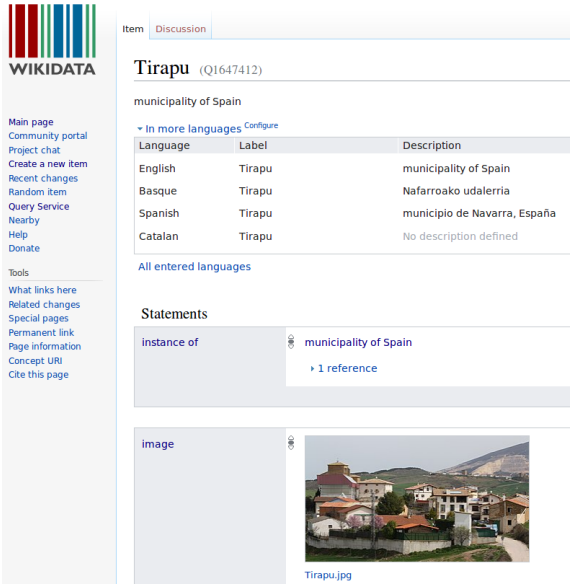

Figure 3: *Tirapu* concept in Wikidata.

different kinds of image (Person, Human-Group, Place/Location, Institution, Building, Animal/Plant, Event/Sport, History, Map/Icon, Culture, and Others). Our hypothesis was that these tags could be almost entirely automatically inferred from their Wikimedia context. A priori, extracting this new information from Wikimedia seemed easy: inferring the tag from the category assigned to the image in Wikimedia Commons. However, the task ultimately proved to be more difficult than expected. The problem is that Commons category system is a folk taxonomy. Not every relation between a category and its parent is an 'is-a' relationship. For example, to know whether an image corresponds to a person a query to Petscan asking for those images whose category in Commons (with a maximum depth of 5) could be "People" ([Query PSID 4346397]), but unfortunately many of the 1,492 results are not those of persons. Using a maximum depth of 3 ([Query PSID 4346796]) deviation is not so wide, the results get only 248 results, but again many of them are not persons.

The alternative was using Wikidata instead, as its taxonomy is more reliable. We defined two reliable automatic ways to extract that information: (1) directly from Wikidata, and (2) indirectly via the Wikipedia pages where the image is used.

Directly from Wikidata: If the image is the one used to illustrate a concept in Wikidata, the desired semantic tag will be the value of the property "instance-of" for that concept in Wikidata. The coverage is 5.5% (196 images are tagged over 3,560 captions). For example, in Figure 3

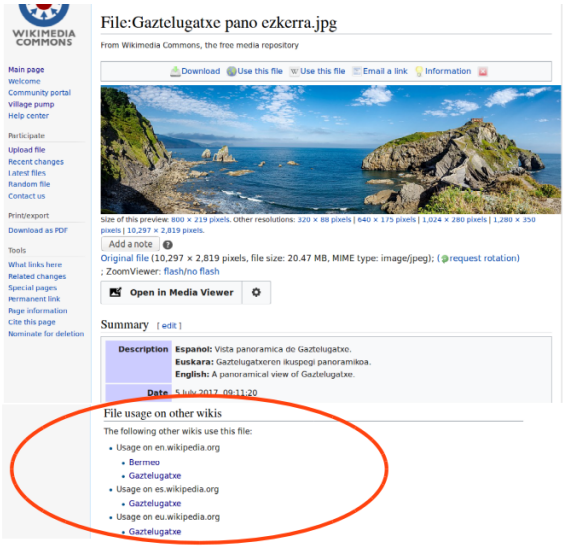

Figure 4: Image usage on other wikis.

the caption of the image Tirapu.jpg is tagged as *Institution* as it is the image for the concept Q1647412, defined in Wikidata as an instance of "municipality of Spain" (institution).

Indirectly via the Wikipedia pages where the image is used: Firstly, given one image, we obtain all the titles of Wikipedia pages in Basque, Spanish and English where the image is used (Wikimedia Commons offers this information). Secondly, we extract the Wikidata identifier for each of those Wikipedia pages. Thirdly, among those concepts we select the one repeated in most languages (Basque, Spanish and/or English). Finally, we extract the "instance-of" value in Wikidata. For example, the image *Gaztelugatxe_pano_ezkerra.jpg* is not directly related to any concept in Wikidata, but looking at its "*File usage on other wikis*" information (see Figure 3.2), we can see that it is used in *Gaztelugatxe* and *Bermeo* articles on the English Wikipedia (Q1496690 and Q695444 concepts in Wikidata).

As the concept Q1496690 is the most representative, using "island" (its value for the "instance-of" property) we assign the tag "Place/Loc". The coverage of this second way of tagging is 45.6% of the images (1,623 over 3,560).

Implicit information in the name of the image: Complementarily, 33.4% of the tags could be obtained looking for 40 words inside the name of the image. For example, as the word 'ermita' ('hermitage' in English) is in the name of the image named *San_Esteban_ermita.jpg*, the image is tagged as 'Building'. The following are other alternative words to tag buildings: hotel, palace, alcázar, museo (museum), tomb, teatro (theatre), zubi (bridge), puente (bridge), architecture, eliza (church), ermita (hermitage) and geltokia (station). After accumulating the explained automatic criteria, a total of 64.9% of the images were tagged automatically, and therefore only 35.1% had to be selected by hand. Our *Tagged-CommonsCaptions* corpus is also publicly available online[6].

4 *Design and evaluation of the MT system*

Following the description of the parallel corpora related to the domain, and the current state-of-the-art MT technology, in this section we describe the method used to build the MT engines (baseline, domain-adapted system and semantically enriched system) and their evaluation.

4.1 Baseline system

The system created to be the state-of-the-art baseline for our experiments uses the optimal configuration for generic MT in Basque-Spanish, recently determined by Etchegoyhen et al. (2018). The NMT systems used in the experiments have been built using the OpenNMT-py toolkit (Klein et al., 2017). The models follow the attention-based encoder-decoder approach (Bahdanau, Cho, and Bengio, 2014). The encoder and the decoder consist of a 4-layer Recurrent Neural Network (RNN) with 800 LSTM (Hochreiter and Schmidhuber, 1997) hidden units, using a vocabulary size of 50,000 in each language. The models were trained for 13 epochs using Stochastic Gradient Descent with an initial learning rate of 1 and applying a learning decay of 0.7 after epoch 10.

The results of Etchegoyhen et al. (2018) show that the use of morphological processing is time-consuming and does not achieve a large improvement. Taking that into account, we decided not to perform any linguistically-motivated approach to segmentation. Hence, the pre-processing of the data is language-independent. The sentences were tokenised and truecased (proper capitalization recovered), and we applied Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch, 2016) (trained in both languages using 30,000 merge operations).

In Table 1 we find a summary of the datasets. To build representative translation models for the Basque-Spanish language pair the Elhuyar Corpus was used. It contains 2 million sentences

[6] http://ixa.si.ehu.eus/node/11513?language=en

	Elhuyar	Captions
train	2,061,863	3,000
dev	3,482	–
test	3,405	560

Table 1: Number of parallel sentences in the eu-es Elhuyar and *CommonsCaptions* datasets.

for both languages, collected from professional translations in different domains, and bilingual web pages (*train* row in Table 1). This is a reduced version of the corpus used by Etchegoyhen et al. (2018), where they also included an extra set of 807,222 sentences collected from comparable data in the news domain.

For building the models adapted for the image captions, we use the *CommonsCaptions* corpus (explained in Section 3.1). We use 3,000 sentences for training and the remaining 560 sentences as our *test* set. The development (*dev*) set used for the models was the same as the Elhuyar *dev* set.

		Elhuyar test	Captions test
es → eu	BLEU	17.61	19.68
	NIST	5.80	5.17
	TER	65.74	69.09
eu → es	BLEU	27.58	23.91
	NIST	7.52	4.94
	TER	56.69	60.27

Table 2: Results of the BLM models.

In Table 2 we provide the results of the baselines, (*BLM models*) using several evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and TER (Snover et al., 2006). These scores indicate how good the outputs of the NMT systems are compared to a human-translated reference. For BLEU and NIST, the higher the score, the better the translation is estimated to be, while for TER goes, as it is an error rate, the lower the better.

4.2 Domain-adapted NMT system

The domain adaptation using the *CommonsCaptions* corpus was designed following the fine-tuning approach (Luong and Manning, 2015; Poncelas, de Buy Wenniger, and Way, 2018b).

In these experiments, the models explained in Section 4.1 are fine-tuned with the *CommonsCaptions* corpus. The models were trained for 12 epochs with the general Elhuyar corpus, with the last iteration trained only with the sentences from the domain corpus (*CommonsCaptions*). In total there were 13 epochs, the same as for the baseline models.

Table 3 summarises the results of the evaluation of domain-adapted models (*ADPM models*). We can see that the in-domain system shows a big improvement with respect to the baseline *BLM* system in terms of BLEU and TER metrics. The improvement in BLEU of ADPM models with respect to BLM (Table 2) when evaluated in *CommonsCaptions* test ranges from 19.68 to 23.01 (relative improvement of 18%) for the es-eu pair and from 23.91 to 29.59 (relative improvement of 23%) for the alternative direction. TER also decreases (meaning an increment of quality) from 69.09 to 62.61 (relative improvement of 9%) for the es-eu pair and 60.27 to 54.08 (relative improvement of 10%) for the eu-es pair.

When using the NIST metric, we see a small drop in performance. The proposed reason for that is that *n*-gram statistics substantially differ in the general corpus and in the domain of captions. While BLEU simply calculates *n*-gram precision adding equal weight to each one, NIST also calculates how informative a particular *n*-gram is. That is to say, when a correct *n*-gram is found, the rarer that *n*-gram is, the more weight it will be given.

		Elhuyar test	Captions test
es → eu	BLEU	15.06	23.01
	NIST	4.94	4.92
	TER	67.92	62.61
eu → es	BLEU	25.41	29.59
	NIST	7.10	5.40
	TER	59.22	54.08

Table 3: Results of the ADPM models.

4.3 Domain-adapted and "semantic sensitive" NMT system

A new experiment was designed to examine whether the rich semantic knowledge inherent to the Wikipedia context can improve the results.

The new system is exactly the same as the previous system described in Section 4.2, but the corpus is replaced with a new version that includes a "semantic tag" as complementary and separated information in both languages. For example, the parallel texts for the image Basurto_ospitalea_geltokia_1.jpg in the *CommonsCaptions* corpus were the following:
eu: *Basurtu-Ospitalea tren geltokia*
es: *Estación de tren Basurto-Hospital*

		Captions test (categorised)
es → eu	BLEU	23.26
	NIST	5.12
	TER	61.63
eu → es	BLEU	29.71
	NIST	5.58
	TER	53.65

Table 4: Results of the SMM models.

		BLM	ADPM	SMM
es → eu	BLEU	19.68	23.01	23.26
	NIST	5.17	4.92	5.12
	TER	69.09	62.61	61.63
eu → es	BLEU	23.91	29.59	29.71
	NIST	4.94	5.40	5.58
	TER	60.27	54.04	53.65

Table 5: Results of the 3 models (BLM, ADPM and SMM) using Captions test for the es-eu pair.

and now the parallel texts in the *TaggedCommonsCaptions* corpus are the following:
eu: *Building ||| Basurtu-Ospitalea tren geltokia*
es: *Building ||| Estación de tren Basurto-Hospital*. As in Section 4.2, the new semantic sensitive models (*SMM models*) is trained for 12 epochs with the general Elhuyar corpus, and then once more with the tagged domain corpus in the 13th iteration.

Table 4 summarises the results of the evaluation of the eu-es and es-eu systems. Figures show an improvement in the three metrics, but unfortunately none of them are statistically significant (using Bootstrap Resampling (Koehn, 2004) at level p=0.01). Table 5 shows a summary of the results of the three systems when they are tested on the *CommonsCaptions* test. The improvement in BLEU ranges from 19.68 to 23.26 (relative improvement of 18%) for the es-eu pair and from 23.91 to 29.71 (relative improvement of 24%) for the alternative direction.

5 Experiments with other language pairs

As the methodology is language-independent we repeat the experiment for the English-Irish (en-ga) pair. We knew that the obtained corpus would be smaller (7,554 words), but taking into account the poor performance in this domain of the baseline en-ga bilingual corpus we wanted to test the contribution of the supplementary corpus.

We present a summary of the data used in Table 6. As baseline data we use a collection of parallel sentences in the public administration domain from diverse sources crawled from the Web [7] and provided by different organizations: Department of Culture Heritage and the Gaeltacht (DCHG), Digital Corpus of the European Parliament (DCEP), Directorate General for Translation (DGT-TM) and Conradh na Gaeilge (CnaG).

Using Petscan to retrieve images with captions in both English and Irish we obtained 2,738 images captions [8]. After performing a cleaning process similar to that described in Section 3.1 for the es-eu pair, the remaining 434 captions were included in the final corpus: 350 were used for training and 84 for testing as shown in Table 6.

	Irish data	Captions
train	108,796	350
dev	655	–
test	1,516	84

Table 6: Number of parallel sentences in the Irish and *CommonsCaptions* datasets.

In order to ensure that this work would be comparable to other studies on en-ga MT, we chose to use the same corpus used by Dowling et al. (2018) in their NMT experiments mentioned in Section 2. This data contains a mixture of crawled data, data from an Irish government department, data from EU institutions, as well as a small amount of data from other sources.

The configuration of the NMT models trained in the en-ga language pairs is the same as that used in en-eu models. In Table 7 we present the results of the models trained on the baseline data and the fine-tuned model. The results show that significant improvements (with $p = 0.01$) are also possible for this pair when using the caption corpus [9], even when the number of sentences is 10 times smaller.

6 Conclusions and Future Work

We have created three corpora (Spanish-Basque *CommonsCaptions*, Spanish-Basque *TaggedCommonsCaptions* and English–Irish *CommonsCaptions*) that are considerable contributions for two low-resource languages. These corpora have been extracted from open data in

[7] `http://www.citizensinformation.ie` and `https://www.teagasc.ie/` websites

[8] `https://petscan.wmflabs.org/?psid=4414745`

[9] `http://ixa.si.ehu.eus/node/11513?language=en`

		BLM	ADPM
en → ga	BLEU	15.24	19.61
	NIST	2.80	2.71
	TER	72.97	68.74
ga → en	BLEU	10.18	14.84
	NIST	2.36	3.05
	TER	74.02	66.86

Table 7: Results of the baseline and the adapted systems evaluated with *CommonsCaptions* test for the en-ga language pair.

Wikimedia. We have also demonstrated that a simple adaptation of a general NMT system to this domain produces significant improvements in translation quality. Supplementary small improvements were detected when using additional semantic tags extracted from the Wikipedia context where the images are being used. The new method to create bilingual corpora and to use it to adapt a general NMT system is language-independent and could be useful to supply Wikimedia Commons content in more languages in the future. We have tested it for the Spanish-Basque language pair and confirmed its utility in the English-Irish pair.

There is broad scope for improving these results, especially when taking into account that when we created our *CommonsCaptions* corpus we collected captions for 3,560 images, but 19,971 images with captions in Basque but not in Spanish were left out. In addition to this there are more than one million images with captions in Spanish but not in Basque.

Another possible improvement is the creation of bigger corpora with additional text information from Wikipedia and Wikidata: (1) the list of the titles in Basque and Spanish for all the articles in Wikipedia [10]; (2) descriptions in Wikidata for both languages; or (3) the first sentence in the contents in Basque and Spanish for all the articles in Wikipedia (if the number of words is not very different).

Acknowledgments

We are indebted to Elhuyar Foundation for providing the corpus to be used in this research. The research leading to these results was carried out as part of the TADEEP project (Spanish Ministry of Economy and Competitiveness TIN2015-70214-P, with FEDER funding). This work has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

[10] Wikipedia_titles_es_eu_2018 Corpus: `http://ixa.si.ehu.eus/node/11500?language=en`

References

[Alegria et al.2013] Alegria, I., U. Cabezon, U. F. de Betono, G. Labaka, A. Mayor, K. Sarasola, and A. Zubiaga. 2013. Reciprocal enrichment between basque wikipedia and machine translation. In *The People's Web Meets NLP*. Springer, pages 101–118.

[Bahdanau, Cho, and Bengio2014] Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

[Bojar et al.2017] Bojar, O., R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark.

[Crego et al.2016] Crego, J., J. Kim, G. Klein, A. Rebollo, K. Yang, J. Senellart, E. Akhanov, P. Brunelle, A. Coquard, Y. Deng, et al. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

[Doddington2002] Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Diego, CA.

[Dowling et al.2015] Dowling, M., L. Cassidy, E. Maguire, T. Lynn, A. Srivastava, and J. Judge. 2015. Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. In *Proceedings of the The Fourth LRL Workshop: Language Technologies in support of Less-Resourced Languages*, pages 314–318, Poznan, Poland.

[Dowling et al.2018] Dowling, M., T. Lynn, A. Poncelas, and A. Way. 2018. Smt versus NMT: Preliminary comparisons for Irish. In *Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20, Boston, USA.

[Etchegoyhen et al.2018] Etchegoyhen, T., E. M. Garcia, A. Azpeitia, G. Labaka, I. Alegria,

I. C. Etxabe, A. J. Carrera, I. E. Santos, and M. M. eta Eusebi Calonge. 2018. Neural machine translation of Basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT)*, pages 139–148, Alicante, Spain.

[Hochreiter and Schmidhuber1997] Hochreiter, S. and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–1780.

[Klein et al.2017] Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada.

[Koehn2004] Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

[Labaka, Alegria, and Sarasola2016] Labaka, G., I. Alegria, and K. Sarasola. 2016. Domain adaptation in MT using titles in wikipedia as a parallel corpus: Resources and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2209–2213, Portorož, Slovenia.

[Labaka et al.2014] Labaka, G., C. España-Bonet, L. Màrquez, and K. Sarasola. 2014. A hybrid machine translation architecture guided by syntax. *Machine translation*, 28(2):91–125.

[Luong and Manning2015] Luong, M.-T. and C. D. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79, Da Nang, Vietnam.

[Nothman et al.2013] Nothman, J., N. Ringland, W. Radford, T. Murphy, and J. R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

[Otero and López2010] Otero, P. G. and I. G. López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pages 21–25, Valletta, Malta.

[Papineni et al.2002] Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

[Poncelas, de Buy Wenniger, and Way2018a] Poncelas, A., G. M. de Buy Wenniger, and A. Way. 2018a. Data selection with feature decay algorithms using an approximated target side. In *15th International Workshop on Spoken Language Translation*, pages 173–180, Bruges, Belgium.

[Poncelas, de Buy Wenniger, and Way2018b] Poncelas, A., G. M. de Buy Wenniger, and A. Way. 2018b. Feature decay algorithms for neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 239–248, Alacant, Spain.

[Poncelas, Way, and Sarasola2018] Poncelas, A., A. Way, and K. Sarasola. 2018. The adapt system description for the iwslt 2018 basque to english translation task. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 76–82, Bruges, Belgium.

[Sennrich, Haddow, and Birch2016] Sennrich, R., B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725, Berlin, Germany,.

[Snover et al.2006] Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

[Wu et al.2016] Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.