A Factory of Comparable Corpora from Wikipedia

Alberto Barrón-Cedeño[1], **Cristina España-Bonet**[2], **Josu Boldoba**[2] and **Lluís Màrquez**[1]
[1] Qatar Computing Research Institute, HBKU, Doha, Qatar
[2] TALP Research Center, Univesitat Politècnica de Catalunya, Barcelona, Spain
{albarron,lmarquez}@qf.org.qa
cristinae@cs.upc.edu jboldoba08@gmail.com

Abstract

Multiple approaches to grab comparable data from the Web have been developed up to date. Nevertheless, coming out with a high-quality comparable corpus of a specific topic is not straightforward. We present a model for the automatic extraction of comparable texts in multiple languages and on specific topics from Wikipedia. In order to prove the value of the model, we automatically extract parallel sentences from the comparable collections and use them to train statistical machine translation engines for specific domains. Our experiments on the English–Spanish pair in the domains of Computer Science, Science, and Sports show that our in-domain translator performs significantly better than a generic one when translating in-domain Wikipedia articles. Moreover, we show that these corpora can help when translating out-of-domain texts.

1 Introduction

Multilingual corpora with different levels of comparability are useful for a range of natural language processing (NLP) tasks. Comparable corpora were first used for extracting parallel lexicons (Rapp, 1995; Fung, 1995). Later they were used for feeding statistical machine translation (SMT) systems (Uszkoreit et al., 2010) and in multilingual retrieval models (Schönhofen et al., 2007; Potthast et al., 2008). SMT systems estimate the statistical models from bilingual texts (Koehn, 2010). Since only the words that appear in the corpus can be translated, having a corpus of the right domain is important to have high coverage. However, it is evident that no large collections of parallel texts for all domains and language pairs exist. In some cases, only general-domain parallel corpora are available; in some others there are no parallel resources at all.

One of the main sources of parallel data is the Web: websites in multiple languages are crawled and contents retrieved to obtain multilingual data. Wikipedia, an on-line community-curated encyclopædia with editions in multiple languages, has been used as a source of data for these purposes —for instance, (Adafre and de Rijke, 2006; Potthast et al., 2008; Otero and López, 2010; Plamada and Volk, 2012). Due to its encyclopædic nature, editors aim at organising its content within a dense taxonomy of categories.[1] Such a taxonomy can be exploited to extract comparable and parallel corpora on specific topics and knowledge domains. This allows to study how different topics are analysed in different languages, extract multilingual lexicons, or train specialised machine translation systems, just to mention some instances. Nevertheless, the process is not straightforward. The community-generated nature of the Wikipedia has produced a reasonably good —yet chaotic— taxonomy in which categories are linked to each other at will, even if sometimes no relationship among them exists, and the borders dividing different areas are far from being clearly defined.

The rest of the paper is distributed as follows. We briefly overview the definition of comparability levels in the literature and show the difficulties inherent to extracting comparable corpora from Wikipedia (Section 2). We propose a simple and effective platform for the extraction of comparable corpora from Wikipedia (Section 3). We describe a simple model for the extraction of parallel sentences from comparable corpora (Section 4). Experimental results are reported on each of these sub-tasks for three domains using the English and Spanish Wikipedia editions. We present an application-oriented evaluation of the comparable corpora by studying the impact of the extracted parallel sentences on a statistical machine translation system (Section 5). Finally, we draw conclusions and outline ongoing work (Section 6).

[1] http://en.wikipedia.org/wiki/Help:Category

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 3–13,
Beijing, China, July 30, 2015.

2 Background

Comparability in multilingual corpora is a fuzzy concept that has received alternative definitions without reaching an overall consensus (Rapp, 1995; Eagles Document Eag–Tcwg–Ctyp, 1996; Fung, 1998; Fung and Cheung, 2004; Wu and Fung, 2005; McEnery and Xiao, 2007; Sharoff et al., 2013). Ideally, a comparable corpus should contain texts in multiple languages which are similar in terms of *form* and *content*. Regarding content, they should observe similar structure, function, and a long list of characteristics: register, field, tenor, mode, time, and dialect (Maia, 2003).

Nevertheless, finding these characteristics in real-life data collections is virtually impossible. Therefore, we attach to the following simpler four-class classification (Skadiņa et al., 2010): (*i*) *Parallel texts* are true and accurate translations or approximate translations with minor language-specific variations. (*ii*) *Strongly comparable texts* are closely related texts reporting the same event or describing the same subject. (*iii*) *Weakly comparable texts* include texts in the same narrow subject domain and genre, but describing different events, as well as texts within the same broader domain and genre, but varying in sub-domains and specific genres. (*iv*) *Non-comparable texts* are pairs of texts drawn at random from a pair of very large collections of texts in two or more languages.

Wikipedia is a particularly suitable source of multilingual text with different levels of comparability, given that it covers a large amount of languages and topics.[2] Articles can be connected via interlanguage links (i.e., a link from a page in one Wikipedia language to an *equivalent* page in another language). Although there are some missing links and an article can be linked by two or more articles from the same language (Hecht and Gergle, 2010), the number of available links allows to exploit the multilinguality of Wikipedia.

Still, extracting a comparable corpus on a specific domain from Wikipedia is not so straightforward. One can take advantage of the user-generated categories associated to most articles. Ideally, the categories and sub-categories would compose a hierarchically organized taxonomy, e.g., in the form of a category tree. Nevertheless,

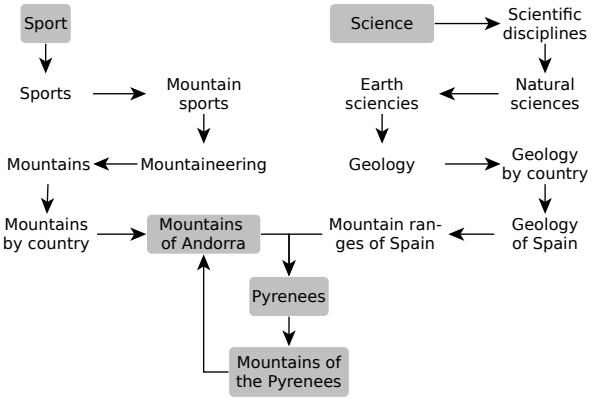

Figure 1: Slice of the Spanish Wikipedia category graph (as in May 2015) departing from categories `Sport` and `Science`. Translated for clarity.

the categories in Wikipedia compose a densely-connected graph with highly overlapping categories, cycles, etc. As they are manually-crafted, the categories are somehow arbitrary and, among other consequences, the potential categorisation of articles does not accomplish with the properties for representing the desirable —trusty enough— categorisation of articles from different domains. Moreover, many articles are not associated to the categories they should belong to and there is a phenomenon of over-categorization.[3]

Figure 1 is an example of the complexity of Wikipedia's category graph topology. Although this particular example comes from the Wikipedia in Spanish, similar phenomena exist in other editions. Firstly, the paths from different *apparently* unrelated categories —`Sport` and `Science`—, converge in a common node soon in the graph (node `Pyrenees`). As a result, not only `Pyrenees` could be considered as a sub-category of both `Sport` and `Science`, but all its descendants. Secondly, cycles exist among the different categories, as in the sequence `Mountains of Andorra` → `Pyrenees` → `Mountains of the Pyrenees` → `Mountains of Andorra`. Ideally, every sub-category of a category should share the same attributes, since the "failure to observe this principle reduces the predictability [of the taxonomy] and can lead to cross-classification" (Rowley and Hartley, 2000, p. 196). Although fixing this issue —inherent to all the Wikipedia editions— falls

[2] Wikipedia contains 288 language editions out of which 277 are active and 12 have more than 1M articles at the time of writing, June 2015 (`http://en.wikipedia.org/wiki/List_of_Wikipedias`).

[3] This is a phenomenon specially stressed in the Wikipedia itself: `http://en.wikipedia.org/wiki/Wikipedia:Overcategorization`.

out of the scope of our research, some heuristic strategies are necessary to diminish their impact in the domain definition process.

Plamada and Volk (2012) dodge this issue by extracting a domain comparable corpus using IR techniques. They use the characteristic vocabulary of the domain (100 terms extracted from an external in-domain corpus) to query a Lucene search engine[4] over the whole encyclopædia. Our approach is completely different: we try to get along with Wikipedia's structure with a strategy to walk through the category graph departing from a root or *pseudo-root* category, which defines our domain of interest. We empirically set a threshold to stop exploring the graph such that the included categories most likely represent an entire domain (cf. Section 3). This approach is more similar to Cui et al. (2008), who explore the *Wiki-Graph* and score every category in order to assess its likelihood of belonging to the domain.

Other tools are being developed to extract corpora from Wikipedia. Linguatools[5] released a comparable corpus extracted from Wikipedias in 253 language pairs. Unfortunately, neither their tool nor the applied methodology description are available. CatScan2[6] is a tool that allows to explore and search categories recursively. The Accurat toolkit (Pinnis et al., 2012; Ştefănescu, Dan and Ion, Radu and Hunsicker, Sabine, 2012)[7] aligns comparable documents and extracts parallel sentences, lexicons, and named entities. Finally, the most related tool to ours: CorpusPedia[8] extracts non-aligned, softly-aligned, and strongly-aligned comparable corpora from Wikipedia (Otero and López, 2010). The difference with respect to our model is that they only consider the articles associated to one specific category and not to an entire domain.

The inter-connection among Wikipedia editions in different languages has been exploited for multiple tasks including lexicon induction (Erdmann et al., 2008), extraction of bilingual dictionaries (Yu and Tsujii, 2009), and identification of particular translations (Chu et al., 2014; Prochasson and Fung, 2011). Different cross-language NLP tasks have particularly taken advantage of Wikipedia. Articles have been used for query translation (Schönhofen et al., 2007) and cross-language semantic representations for similarity estimation (Cimiano et al., 2009; Potthast et al., 2008; Sorg and Cimiano, 2012). The extraction of parallel corpora from Wikipedia has been a hot topic during the last years (Adafre and de Rijke, 2006; Patry and Langlais, 2011; Plamada and Volk, 2012; Smith et al., 2010; Tomás et al., 2008; Yasuda and Sumita, 2008).

3 Domain-Specific Comparable Corpora Extraction

In this section we describe our proposal to extract domain-specific comparable corpora from Wikipedia. The input to the pipeline is the top category of the domain (e.g., `Sport`). The terminology used in this description is as follows. Let c be a Wikipedia category and c^* be the top category of a domain. Let a be a Wikipedia article; $a \in c$ if a contains c among its categories. Let G be the Wikipedia category graph.

Vocabulary definition. The domain vocabulary represents the set of terms that better characterises the domain. We do not expect to have at our disposal the vocabulary associated to every category. Therefore, we build it from the Wikipedia itself. We collect every article $a \in c^*$ and apply standard pre-processing; i.e., tokenisation, stopwording, numbers and punctuation marks filtering, and stemming (Porter, 1980). In order to reduce noise, tokens shorter than four characters are discarded as well. The vocabulary is then composed of the top n terms, ranked by term frequency. This value is empirically determined.

Graph exploration. The input for this step is G, c^* (i.e., the departing node in the graph), and the domain vocabulary. Departing from c^*, we perform a breadth-first search, looking for all those categories which more likely belong to the required domain. Two constraints are applied in order to make a controlled exploration of the graph: (*i*) in order to avoid loops and exploring already traversed paths, a node can only be visited once, (*ii*) in order to avoid exploring the whole categories graph, a stopping criterion is pre-defined. Our stopping criterion is inspired by the classification tree-breadth first search algorithm (Cui et al., 2008). The core idea is scoring the explored cate-

[4] `https://lucene.apache.org/`
[5] `http://linguatools.org`
[6] `http://tools.wmflabs.org/catscan2/catscan2.php`
[7] `http://www.accurat-project.eu`
[8] `http://gramatica.usc.es/pln/tools/CorpusPedia.html`

Edition	Articles	Categories	Ratio
English	4,123,676	1,032,222	4.0
Spanish	965,543	210,803	4.6
Intersection	631,710	107,313	–

Table 1: Amount of articles and categories in the Wikipedia editions and in the intersection (i.e., pages linked across languages).

gories to determine if they belong to the domain. Our heuristic assumes that a category belongs to the domain if its title contains at least one of the terms in the characteristic vocabulary. Nevertheless, many categories exist that may not include any of the terms in the vocabulary. (e.g., consider category `pato` in Spanish —literally "duck" in English— which, somehow surprisingly, refers to a sport rather than an animal). Our naïve solution to this issue is to consider subsets of categories according to their depth respect to the root. An entire level of categories is considered part of the domain if a minimum percentage of its elements include vocabulary terms.

In our experiments we use the English and Spanish Wikipedia editions.[9] Table 1 shows some statistics, after filtering disambiguation and redirect pages. The intersection of articles and categories between the two languages represents the ceiling for the amount of parallel corpora one can gather for this pair. We focus on three domains: Computer Science (CS), Science (Sc), and Sports (Sp) —the top categories c^* from which the graph is explored in order to extract the corresponding comparable corpora.

Table 2 shows the number of *root articles* associated to c^* for each domain and language. From them, we obtain domain vocabularies with a size between 100 and 400 lemmas (right-side columns) when using the top 10% terms. We ran experiments using the top 10%, 15%, 20% and 100%. The relatively small size of these vocabularies allows to manually check that 10% is the best option to characterise the desired category, higher percentages add more noise than in-domain terms. The plots in Figure 2 show the percentage of categories with at least one domain term in the title: the starting point for our graph-based method for selecting the in-domain articles. As expected, nearly 100% of the categories in the root include domain terms and this percentage decreases with increasing depth in the tree.

[9]Dumps downloaded from `https://dumps.wikimedia.org` in July 2013 and pre-processed with JWPL (Zesch et al., 2008) (`https://code.google.com/p/jwpl/`).

	Articles		Vocabulary	
	en	es	en	es
CS	4	130	106	447
Sc	29	3	464	140
Sp	3	10	122	100

Table 2: Number of articles in the root categories and size of the resulting domain vocabulary.

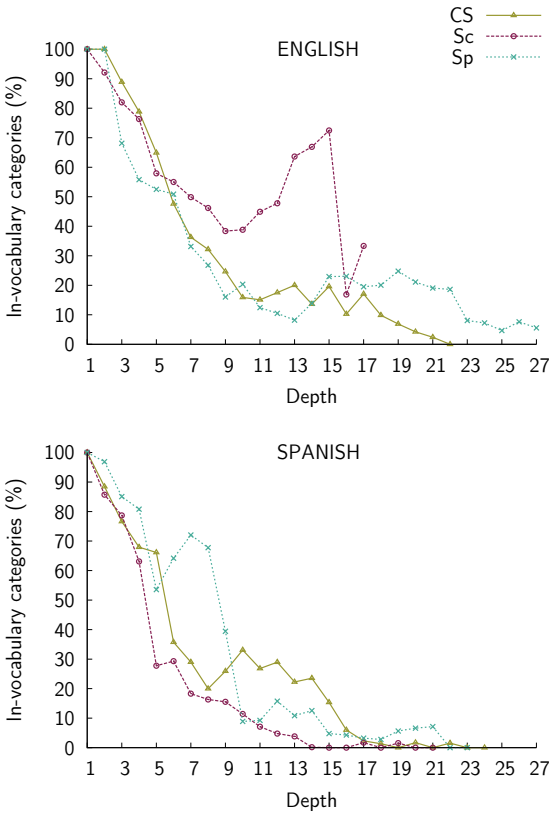

Figure 2: Percentage of categories with at least one domain term in the title for the two languages and the three domains under study.

When extracting the corpus, one must decide the adequate percentage of positive categories allowed. High thresholds lead to small corpora whereas low thresholds lead to larger —but noisier— corpora. As in many applications, this is a trade-off between precision and recall and depends on the intended use of the corpus. Table 3 shows some numbers on two different thresholds. Increasing the threshold does not always mean

	Articles		Distance from the root			
	50%	60%	50%		60%	
	en-es	en-es	en	es	en	es
CS	18,168	8,251	6	5	5	5
Sc	161,130	21,459	6	4	4	4
Sp	72,315	1,980	8	8	3	4

Table 3: Number of article pairs according to the percentage of positive categories used to select the levels of the graph and distance from the root at which the percentage is smaller to the desired one.

lowering the selected depth, but when it does, the difference in the number of extracted articles can be significant. The same table shows the number of article pairs extracted for each value: the resulting comparable corpus for each domain. The stopping level is selected for every language independently, but in order to reduce noise, the comparable corpus is only built from those articles that appear in both languages and are related via an interlanguage link. We validate the quality in terms of application-based utility of the generated comparable corpora when used in a translation system (cf. Section 5). Therefore, we choose to give more importance to recall and opt for the corpora obtained with a threshold of 50%.

4 Parallel Sentence Extraction

In this section we describe a simple technique for extracting parallel sentences from a comparable corpus.

Given a pair of articles related by an interlanguage link, we estimate the similarity between all their pairs of cross-language sentences with different text similarity measures. We repeat the process for all the pairs of articles and rank the resulting sentence pairs according to its similarity. After defining a threshold for each measure, those sentence pairs with a similarity higher than the threshold are extracted as parallel sentences. This is a non-supervised method that generates a noisy parallel corpus. The quality of the similarity measures will then affect the purity of the parallel corpus and, therefore, the quality of the translator. However, we do not need to be very restrictive with the measures here and still favour a large corpus, since the word alignment process in the SMT system can take care of part of the noise.

Similarity computation. We compute similarities between pairs of sentences by means of cosine and length factor measures. The cosine similarity is calculated on three well-known characterisations in cross-language information retrieval and parallel corpora alignment: (*i*) character n-grams (cng) (McNamee and Mayfield, 2004); (*ii*) pseudo-cognates (cog) (Simard et al., 1992); and (*iii*) word 1-grams, after translation into a common language, both from English to Spanish and vice versa (mono_{en}, mono_{es}). We add the (*iv*) length factor (len) (Pouliquen et al., 2003) as an independent measure and as penalty (multiplicative factor) on the cosine similarity.

The threshold for each of the measures just introduced is empirically set in a manually annotated corpus. We define it as the value that maximises the F_1 score on this development set. To create this set, we manually annotated a corpus with 30 article pairs (10 per domain) at sentence level. We considered three sentence classes: parallel, comparable, and other. The volunteers of the exercise were given as guidelines the definitions by Skadiņa et al. (2010) of *parallel text* and *strongly comparable text* (cf. Section 2). A pair that did not match any of these definitions had to be classified as other. Each article pair was annotated by two volunteers, native speakers of Spanish with high command of English (a total of nine volunteers participated in the process). The mean agreement between annotators had a kappa coefficient (Cohen, 1960) of $\kappa \sim 0.7$. A third annotator resolved disagreed sentences.[10]

Table 4 shows the thresholds that obtain the maximum F_1 scores. It is worth noting that, even if the values of precision and recall are relatively low —the maximum recall is 0.57 for len—, our intention with these simple measures is not to obtain the highest performance in terms of retrieval, but injecting the most useful data to the translator, even at the cost of some noise. The performance with character 3-grams is the best one, comparable to that of mono, with an F_1 of 0.36. This suggests that a translator is not mandatory for performing the sentences selection. Len and 1-grams have no discriminating power and lead to the worse scores (F_1 of 0.14 and 0.21, respectively).

We ran a second set of experiments to explore the combination of the measures. Table 5 shows

[10]The corpus is publicly available at `http://www.cs.upc.edu/~cristinae/CV/recursos.php`.

	c1g	c2g	c3g	c4g	c5g	cog	mono$_{en}$	mono$_{es}$	len
Thres.	0.95	0.60	0.25	0.20	0.15	0.30	0.20	0.15	0.90
P	0.18	0.29	0.28	0.24	0.23	0.16	0.30	0.26	0.08
R	0.25	0.31	0.53	0.47	0.47	0.49	0.46	0.34	0.57
F_1	0.21	0.30	0.36	0.32	0.31	0.24	0.36	0.30	0.14

Table 4: Best thresholds and their associated Precision (P), recall (R) and F_1.

	$\bar{S}$	$\bar{S}$·len	$\overline{S \cdot F_1}$	$\overline{S \cdot F_1}$·len
Thres.	0.25	0.15	0.05	0.05
P	0.27	0.33	0.18	0.32
R	0.50	0.62	0.77	0.65
F_1	0.35	0.43	0.29	0.43

Table 5: Precision, recall, and F_1 for the average of the similarities weighted by length model (len) and/or their F_1.

the performance obtained by averaging all the similarities ($\bar{S}$), also after multiplying them by the length factor and/or the observed F_1 obtained in the previous experiment. Even if the length factor had shown a poor performance in isolation, it helps to lift the F_1 figures consistently after affecting the similarities. In this case, F_1 grows up to 0.43. This impact is not so relevant when the individual F_1 is used for weighting $\bar{S}$.

We applied all the measures —both combined and in isolation— on the entire comparable corpora previously extracted. Table 6 shows the amount of parallel sentences extracted by applying the empirically defined thresholds of Tables 4 and 5. As expected, more flexible alternatives, such as low-level n-grams or length factor result in a higher amount of retrieved instances, but in all cases the size of the corpora is remarkable. For the most restricted domain, CS, we get around 200k parallel sentences for a given similarity measure. For the widest domain, SC, we surpass the 1M sentence pairs. As it will be shown in the following section, these sizes are already useful to be used for training SMT systems. Some standard parallel corpora have the same order of magnitude. For tasks other than MT, where the precision on the extracted pairs can be more important than the recall, one can obtain cleaner corpora by using a threshold that maximises precision instead of F_1.

	CS	Sc	Sp
c1g	207,592	1,585,582	404,656
c2g	99,964	745,821	326,882
c3g	96,039	724,210	335,147
c4g	110,701	863,090	394,105
c5g	126,692	1,012,993	466,007
cog	182,981	1,215,008	451,941
len	271,073	1,941,866	550,338
mono$_{en}$	211,209	1,367,917	461,731
mono$_{es}$	183,439	1,273,509	435,671
$\bar{S}$	154,917	1,098,453	450,933
$\bar{S}$·len	121,697	957,662	390,783
$\overline{S \cdot F_1}$	153,056	1,085,502	448,076
$\overline{S \cdot F_1}$·len	121,407	957,967	392,241

Table 6: Size of the parallel corpora extracted with each similarity measure.

5 Evaluation: Statistical Machine Translation Task

In this section we validate the quality of the obtained corpora by studying its impact on statistical machine translation. There are several parallel corpora for the English–Spanish language pair. We select as a general-purpose corpus Europarl v7 (Koehn, 2005), with 1.97M parallel sentences. The order of magnitude is similar to the largest corpus we have extracted from Wikipedia, so we can compare the results in a size-independent way. If our corpus extracted from Wikipedia was made up with parallel fragments of the desired domain, it should be the most adequate to translate these domains. If the quality of the parallel fragments was acceptable, it should also help when translating out-of-domain texts. In order to test these hypotheses we analyse three settings: (*i*) train SMT systems only with Wikipedia (WP) *or* Europarl (EP) to translate domain-specific texts, (*ii*) train SMT systems with Wikipedia *and* Europarl to

translate domain-specific texts, and (*iii*) train SMT systems with Wikipedia *and* Europarl to translate out-of-domain texts (news).

For the out-of-domain evaluation we use the News Commentaries 2011 test set and the News Commentaries 2009 for development.[11] For the in-domain evaluation we build the test and development sets in a semiautomatic way. We depart from the parallel corpora gathered in Section 4 from which sentences with more than four tokens and beginning with a letter are selected. We estimate its perplexity with respect to a language model obtained with Europarl in order to select the most fluent sentences and then we rank the parallel sentences according to their similarity and perplexity. The top-n fragments were manually revised and extracted to build the Wikipedia test (WPtest) and development (WPdev) sets. We repeated the process for the three studied domains and drew 300 parallel fragments for development for every domain and 500 for test. We removed these sentences from the corresponding training corpora. For one of the domains, CS, we also gathered a test set from a parallel corpus of GNOME localisation files (Tiedemann, 2012). Table 7 shows the size in number of sentences of these test sets and of the 20 Wikipedia training sets used for translation. Only one measure, that with the highest F_1 score, is selected from each family: c3g, cog, mono$_{en}$ and $\bar{S}$·len (cf. Tables 4 and 5). We also compile the corpus that results from the union of the previous four. Notice that, although we eliminate duplicates from this corpus, the size of the union is close to the sum of the individual corpora. This indicates that every similarity measure selects a different set of parallel fragments. Beside the specialised corpus for each domain, we build a larger corpus with all the data (Un). Again, duplicate fragments coming from articles belonging to more than one domain are removed.

SMT systems are trained using standard freely available software. We estimate a 5-gram language model using interpolated Kneser–Ney discounting with `SRILM` (Stolcke, 2002). Word alignment is done with `GIZA++` (Och and Ney, 2003) and both phrase extraction and decoding are done with `Moses` (Koehn et al., 2007). We optimise the feature weights of the model with Minimum Error Rate Training (MERT) (Och, 2003) against the BLEU evaluation metric (Papineni et al., 2002). Our model considers the language model, direct and inverse phrase probabilities, direct and inverse lexical probabilities, phrase and word penalties, and a lexicalised reordering.

[11] Both are available at `http://www.statmt.org/wmt14/translation-task.html`.

	CS	Sc	Sp	Un
c3g	95,715	723,760	334,828	883,366
cog	182,283	1,213,965	451,324	1,430,962
mono$_{en}$	210,664	1,367,169	461,237	1,638,777
$\bar{S}$·len	120,835	956,346	389,975	1,160,977
union	577,428	3,847,381	1,181,664	4,948,241
WPdev	300	300	300	900
WPtest	500	500	500	1500
GNOME	1000	–	–	–

Table 7: Number of sentences of the Wikipedia parallel corpora used to train the SMT systems (top rows) and of the sets used for development and test.

	CS	Sc	Sp	Un	Comp.
Europarl	27.99	34.00	30.02	30.63	–
c3g	38.81	40.53	46.94	43.68	43.68
cog	57.32	56.17	57.60	58.14	54.89
mono$_{en}$	54.27	52.96	55.74	55.17	52.45
$\bar{S}$·len	56.14	57.40	58.39	58.80	56.78
union	**64.65**	**62.95**	**62.65**	**64.47**	–

Table 8: BLEU scores obtained on the Wikipedia test sets for the 20 specialised systems described in Section 5. A comparison column (Comp.) where all the systems are trained with corpora of the same size is also included (see text).

(*i*) Training systems with Wikipedia *or* Europarl for domain-specific translation. Table 8 shows the evaluation results on WPtest. All the specialised systems obtain significant improvements with respect to the Europarl system, regardless of their size. For instance, the worst specialised system (c3g with only 95,715 sentences for CS) outperforms by more than 10 points of BLEU the general Europarl translator. The most complete system (the union of the four representatives) doubles the BLEU score for all the domains with an impressive improvement of 30 points. This is of course possible due to the nature of the test set that has been extracted from the same collection as the training data and therefore shares its structure and vocabulary.

To give perspective to these high numbers we evaluate the systems trained on the CS domain

	CS	Un	Comp.
c3g	11.08	9.56	9.56
cog	18.48	17.66	16.31
mono$_{en}$	19.48	20.58	18.84
$\bar{S}$·len	20.71	20.56	19.76
union	**22.41**	20.63	–

Table 9: BLEU scores obtained on the GNOME test set for systems trained only with Wikipedia. A system with Europarl achieves a score of 18.15.

against the GNOME dataset (Table 9). Except for c3g, the Wikipedia translators always outperform the baseline with EP; the union system improves it by 4 BLEU points (22.41 compared to 18.15) with a four times smaller corpus. This confirms that a corpus automatically extracted with an F_1 smaller than 0.5 is still useful for SMT. Notice also that using only the in-domain data (CS) is always better than using the whole WP corpus (Un) even if the former is in general ten times smaller (cf. Table 7).

According to this indirect evaluation of the similarity measures, character n-grams (c3g) represent the worst alternative. These results contradict the direct evaluation, where c3g and mono$_{en}$ had the highest F_1 scores on the development set among the individual similarity measures. The size of the corpus is not relevant here: when we train all the systems with the same amount of data, the ranking in the quality of the measures remains the same. To see this, we trained four additional systems with the top m number of parallel fragments, where m is the size of the smallest corpus for the union of domains: Un-c3g. This new comparison is reported in columns "Comp." in Tables 8 and 9. In this fair comparison c3g is still the worst measure and $\bar{S}$·len the best one. The translator built from its associated corpus outperforms with less than half of the data used for training the general one (883,366 vs. 1,965,734 parallel fragments) both in WPtest (56.78 vs. 30.63) and GNOME (19.76 vs. 18.15).

(*ii*) Training systems on Wikipedia *and* Europarl for domain-specific translation. Now we enrich the general translator with Wikipedia data or, equivalently, complement the Wikipedia translator with out-of-domain data. Table 10 shows the results. Augmenting the size of the in-domain corpus by 2 million fragments improves the results even more, about 2 points of BLEU

	CS	Sc	Sp	Un
Europarl	27.99	34.00	30.02	30.63
union	64.65	62.95	62.65	64.47
EP+c3g	46.07	48.29	50.40	49.34
EP+cog	58.39	57.70	59.05	58.98
EP+mono$_{en}$	54.44	53.93	56.05	55.88
EP+$\bar{S}$·len	56.05	57.53	59.78	58.72
EP+union	**66.22**	**64.24**	**64.39**	**65.67**

Table 10: BLEU scores obtained on the Wikipedia test set for the 20 systems trained with the combination of the Europarl (EP) and the Wikipedia corpora. The results with a Europarl system and the best one from Table 8 (union) shown for comparison.

	CS	Un
EP+c3g	19.78	19.49
EP+cog	21.09	20.14
EP+mono$_{en}$	21.27	20.66
EP+$\bar{S}$·len	21.58	20.65
EP+union	**22.37**	21.43

Table 11: BLEU scores obtained on the GNOME test set for systems trained with Europarl and Wikipedia. A system with Europarl achieves a score of 18.15.

when using all the union data. System c3g benefits the most of the inclusion of the Europarl data. The reason is that it is the individual system with less corpus available and the one obtaining the worst results. In fact, the better the Wikipedia system, the less important the contribution from Europarl is. For the independent test set GNOME, Table 11 shows that the union corpus on CS is better than any combination of Wikipedia and Europarl. Still, as aforementioned, the best performance on this test set is obtained with a pure in-domain system (cf. Table 9).

(*iii*) Training systems on Wikipedia and Europarl for out-of-domain translation. Now we check the performance of the Wikipedia translators on the out-of-domain news test. Table 12 shows the results. In this neutral domain for Europarl and Wikipedia, the in-domain Wikipedia systems show a lower performance. The BLEU score obtained with the Europarl system is 27.02 whereas the Wikipedia union system achieves 22.16. When combining the two corpora, results

	CS	Sc	Sp	Un
union	16.74	22.28	15.82	22.16
EP+c3g	26.06	26.35	26.81	**27.07**
EP+cog	26.61	**27.33**	26.71	**27.08**
EP+mono$_{en}$	**27.18**	26.80	26.96	**27.44**
EP+$\bar{S}$·len	**27.59**	26.80	**27.58**	**27.22**
EP+union	26.76	**27.52**	**27.35**	26.72

Table 12: BLEU scores for the out-of-domain evaluation on the News Commentaries 2011 test set. We show in boldface all the systems that improve the Europarl translator, which achieves a score of 27.02.

are controlled by the Europarl baseline. In general, systems in which we include only texts from an unrelated domain do not improve the performance of the Europarl system alone, results of the combined system are better when we use Wikipedia texts from all the domains together (column Un) for training. This suggests that, as expected, a general Wikipedia corpus is necessary to build a general translator. This is a different problem to deal with.

6 Conclusions and Ongoing Work

In this paper we presented a model for the automatic extraction of in-domain comparable corpora from Wikipedia. It makes possible the automatic extraction of monolingual and comparable article collections as well as a one-click parallel corpus generation for on-demand language pairs and domains. Given a pair of languages and a main category, the model explores the Wikipedia categories graph and identifies a subset of categories (and their associated articles) to generate a document-aligned comparable corpus. The resulting corpus can be exploited for multiple natural language processing tasks. Here we applied it as part of a pipeline for the extraction of domain-specific parallel sentences. These parallel instances allowed for a significant improvement in the machine translation quality when compared to a generic system and applied to a domain specific corpus (in-domain). The experiments are shown for the English–Spanish language pair and the domains Computer Science, Science, and Sports. Still it can be applied to other language pairs and domains.

The prototype is currently operating in other languages. The only prerequisite is the existence of the corresponding Wikipedia edition and some basic processing tools such as a tokeniser and a lemmatiser. Our current efforts intend to generate a more robust model for parallel sentences identification and the design of other indirect evaluation schemes to validate the model performance.

Acknowledgments

This work was partially funded by the TACARDI project (TIN2012-38523-C02) of the Spanish Ministerio de Economía y Competitividad (MEC).

References

Sisay Fissaha Adafre and Maarten de Rijke. 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.

Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors. 2008. *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Iterative Bilingual Lexicon Extraction from Comparable Corpora with Topical and Contextual Knowledge. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 8404 of *Lecture Notes in Computer Science*, pages 296–309. Springer Berlin Heidelberg.

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit Versus Latent Concept Models for Cross-language Information Retrieval. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1513–1518, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gaoying Cui, Qin Lu, Wenjie Li, and Yirong Chen. 2008. Corpus Exploitation from Wikipedia for Ontology Construction. In Calzolari et al. (Calzolari et al., 2008), pages 2126–2128.

Eagles Document Eag–Tcwg–Ctyp. 1996. EAGLES Preliminary recommendations on Corpus Typology.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. An Approach for Extracting Bilingual Terminology from Wikipedia. In

Proceedings of the 13th International Conference on Database Systems for Advanced Applications, DASFAA'08, pages 380–392, Berlin, Heidelberg. Springer-Verlag.

Pascale Fung and Percy Cheung. 2004. Mining verynon-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of EMNLP*, pages 57–63, Barcelona, Spain, July 25–July 26.

Pascale Fung. 1995. Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. In *Proceedings of the Third Annual Workshop on Very Large Corpora*, pages 173–183.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. *Lecture Notes in Computer Science*, 1529:1–17.

Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-generated Content and Its Applications in a Multilingual Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 291–300, New York, NY, USA. ACM.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Machine Translation Summit X*, pages 79–86.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Belinda Maia. 2003. What are comparable corpora. In *Proceedings of the Corpus Linguistics workshop on Multilingual Corpora: Linguistic requirements and technical perspectives*.

Anthony M. McEnery and Zhonghua Xiao, 2007. *Incorporating Corpora: Translation and the Linguist*, chapter Parallel and comparable corpora: What are they up to? Translating Europe. Multilingual Matters.

Paul McNamee and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2):73–97.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51. See also [`http://www.fjoch.com/GIZA++.html`].

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Pablo Gamallo Otero and Issac González López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 21–25, 22 May. Available at http://www.fb06.uni-mainz.de/lk/bucc2010/documents/Proceedings-BUCC-2010.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.

Alexandre Patry and Philippe Langlais. 2011. Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia. In Pierre Zweigenbaum, Reinhard Rapp, and Serge Sharoff, editors, *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95, Portland, Oregon. Association for Computational Linguistics.

Mārcis Pinnis, Radu Ion, Dan Ştefănescu, Fangzhong Su, Inguna Skadiņa, Andrejs Vasiļjevs, and Bogdan Babych. 2012. Accurat toolkit for multi-level alignment and information extraction from comparable corpora. In *Proceedings of the ACL 2012 System Demonstrations*, ACL'12, pages 91–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Magdalena Plamada and Martin Volk. 2012. Towards a Wikipedia-extracted alpine corpus. In *The Fifth Workshop on Building and Using Comparable Corpora*, May.

Martin F. Porter. 1980. An Algorithm for Suffix Stripping. *Program*, 14:130–137.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A Wikipedia-Based Multilingual Retrieval Model. *Advances in Information Retrieval, 30th European Conference on IR Research*, LNCS (4956):522–530. Springer-Verlag.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. 2003. Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 401–408, Borovets, Bulgaria.

Emmanuel Prochasson and Pascale Fung. 2011. Rare Word Translation Extraction from Aligned Comparable Documents. In *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, pages 1327–1335, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying Word Translations in Non-Parallel Texts. *CoRR*, cmp-lg/9505037.

Jennifer Rowley and Richard Hartley. 2000. *Organizing Knowledge. An Introduction to Managing Access to Information*. Ashgate, 3rd edition.

Péter Schönhofen, András A. Benczúr, István Bíró, and Károly Csalogány. 2007. Cross-language retrieval with wikipedia. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 72–79.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum, 2013. *Building and Using Comparable Corpora*, chapter Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. Springer.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Inguna Skadiņa, Ahmet Aker, Voula Giouli, Dan Tufiş, Robert Gaizauskas, Madara Mieriņa, and Nikos Mastropavlos. 2010. A collection of comparable corpora for under-resourced languages. In *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, pages 161–168, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Sorg and Philipp Cimiano. 2012. Exploiting Wikipedia for Cross-lingual and Multilingual Information Retrieval. *Data Knowl. Eng.*, 74:26–45, April.

Ştefănescu, Dan and Ion, Radu and Hunsicker, Sabine. 2012. Hybrid Parallel Sentence Mining from Comparable Corpora. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, Trento, Italy. European Association for Machine Translation .

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling toolkit. In *Intl. Conference on Spoken Language Processing*, Denver, Colorado.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Jesús Tomás, Jordi Bataller, Francisco Casacuberta, and Jaime Lloret. 2008. Mining wikipedia as a parallel and comparable corpus. *LANGUAGE FORUM*, 34(1). Article presented at CICLing-2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics, February 17 to 23, 2008, Haifa, Israel.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In Chu-Ren Huang and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1101–1109, Beijing, China, August. COLING 2010 Organizing Committee.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Natural Language Processing - IJCNLP 2005, Second International Joint Conference*, pages 257–268, Jeju Island, Korea, Oct 11–Oct 13.

Keiji Yasuda and Eiichiro Sumita. 2008. Method for Building Sentence-Aligned Corpus from Wikipedia. In *Association for the Advancement of Artificial Intelligence*.

Kun Yu and Junichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. In *Proceedings of Machine Translation Summit XII*.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wikictionary. In Calzolari et al. (Calzolari et al., 2008).