GeBioToolkit: Automatic Extraction of Gender-Balanced Multilingual Corpus of Wikipedia Biographies

Marta R. Costa-jussà, Pau Li Lin, Cristina España-Bonet*
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
* DFKI GmbH and Saarland University, Saarbrücken
marta.ruiz@upc.edu, lilin.pau@gmail.com, cristinae@dfki.de

Abstract

We introduce GeBioToolkit, a tool for extracting multilingual parallel corpora at sentence level, with document and gender information from Wikipedia biographies. Despite the gender inequalities present in Wikipedia, the toolkit has been designed to extract corpus balanced in gender. While our toolkit is customizable to any number of languages (and to other domains than biographical entries), in this work we present a corpus of 2,000 sentences in English, Spanish and Catalan, which has been post-edited by native speakers to become a high-quality dataset for machine translation evaluation. While GeBioCorpus aims at being one of the first non-synthetic gender-balanced test datasets, GeBioToolkit aims at paving the path to standardize procedures to produce gender-balanced datasets.

Keywords: corpora, gender bias, Wikipedia, machine translation

1. Introduction

Gender biases are present in many natural language processing applications (Costa-jussà, 2019). This comes as an undesired characteristic of deep learning architectures where their outputs seem to reflect demographic asymmetries (Prates et al., 2019). This is of course not due to the architecture itself but to the data used to train a system. Recent research is being devoted to correct the asymmetries mainly by data augmentation techniques in fields such as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018; Webster et al., 2018) or abusive language detection (Park et al., 2018). Test sets have been created in those cases, but we are not aware of any test set available for machine translation (MT).

From another side, machine translation either neural, statistical or rule-based, usually operates in a sentence-by-sentence basis. However, when translating consistently a document, surrounding sentences may have valuable information. The translation of pronouns, verb tenses and even content words might depend on other fragments within a document. This affects also the translation of the gender markers, specially when translating from a language without these marks (e.g. English) into a language with them (e.g. Catalan). Test sets at sentence level are not useful to evaluate these phenomena. But in the time of claims of human parity in MT, all these phenomena are crucial, and document-level evaluation is needed in order to discriminate among systems (Läubli et al., 2018).

Beyond these gender-balanced and document-level needs, the rise of multilingual neural machine translation and the lack of multi-way parallel corpus that can evaluate its abilities (e.g. zero-shot), motivates the creation of new multilingual corpora. In turn, this motivation prompts the development of software to automatically create such corpora.

In order to create a *GeBioToolkit* that is able to systematically extract multilingual parallel corpus at sentence level and with document-level information, we rely on Wikipedia, a free online multilingual encyclopedia written by volunteers. The toolkit is customizable for languages and gender-balance. We take advantage of Wikipedia multilinguality to extract a corpus of biographies, being each biography a document available in all the selected languages. However, given the bias in the representation of males and females also in Wikipedia (Bamman and Smith, 2014) and, to deliver a balanced set, we need to identify and select a subset of documents just after the toolkit performs the extraction of parallel sentences so it can assure parity. Note that our document-level extraction is consistent (all sentences in a document belong to the same personality) but the documents do not keep coherence anymore, since we are removing some sentences —the non-parallel ones— within that document. In our experiments in English–Spanish–Catalan, GeBioToolkit is able to extract parallel sentences with 87.5% accuracy according to a human evaluation.

Besides providing the tool, we also manually post-edit a small subset of the English (en), Spanish (es) and Catalan (ca) outputs to provide a test dataset which is ready-to-use for machine translation applications; to our knowledge the first gender-balanced dataset extracted from real texts in the area: *GeBioCorpus*. With these three languages we aim to cover two linguistic families (Germanic and Romance) which differ in morphology and specifically in gender inflection. The choice in languages is intended to allow the evaluation of machine translation outputs in three different settings: distant morphologies for a high-resourced language pair (English–Spanish) and a low-resourced pair (English–Catalan), and closely related languages (Spanish–Catalan). The latter responds the challenge recently initiated in the WMT International Evaluation[1]. We used native/fluent speakers to provide post-editions on the final multilingual set in the three languages.

The rest of the paper is organized as follows. Section 2. describes some available multilingual datasets used for machine translation evaluation, related work on parallel sen-

[1] http://www.statmt.org/wmt19/similar.html

tence extraction from Wikipedia and a brief mention to general research lines on gender bias in NLP. Section 3. describes the general architecture and Section 4. the methodology, evaluation and characteristics of the extracted corpora. Finally, Section 5. summarizes the work and points at several future extensions of GeBioToolkit.

2. Related Work

There are several multilingual parallel datasets available to evaluate MT outputs. The corpora covering more languages are JRC-Acquis (Acquis Communautaire) and the TED talks corpus. Arab-Acquis (Habash et al., 2017) is a multilingual dataset for 22 European languages plus Arabic with more than 12,000 sentences coming from JRC-Acquis, that is, a collection of legislative texts of the European Union. In order to make the test set equal in all the languages, only sentences that are parallel simultaneously in the 22 languages were extracted (Koehn et al., 2009) and, therefore, the document structure of the data is lost.

The Web Inventory of Transcribed and Translated Talks, WIT^{3}[2], includes English subtitles from TED talks and their translations currently in 109 languages. Parallel corpora are extracted for several pairs (Cettolo et al., 2012) and test sets are annually prepared for the International Workshop on Spoken Language Translation (IWSLT) evaluation campaigns. Test sets exist for all the pairs among German, English, Italian, Dutch and Romanian; and from English to Arabic, Basque, Chinese, Czech, Dutch, Farsi, French, German, Hebrew, Italian, Japanese, Korean, Polish, Portuguese-Brazil, Romanian, Russian, Slovak, Slovenian, Spanish, Thai, Turkish and Vietnamese. In this case the whole talks are aligned at sentence level, so, the document structure is kept but the set is not equivalent in all the languages.

Similarly, the news translation task at the annual workshops and conferences on statistical machine translation (WMT) distributes collections of parallel news to their participants. Test sets have been created over the years for pairs including English into Chinese, Czech, Estonian, Finnish, French, German, Hindi, Hungarian, Kazakh, Latvian, Spanish, Romanian, Russian and Turkish. Again, the document structure is kept in the dataset but, in general, the set is not equivalent in all the languages.

Wikipedia is widely used in natural language processing and it is an excellent resource for multilingual tasks, parallel sentence extraction being among them. However, we do not know of any multilingual corpus of biographies extracted from the resource. On the monolingual side, one can find WikiBiography[3], a corpus of 1,200 biographies in German with automatic annotations of PoS, lemmas, syntactic dependencies, anaphora, discourse connectives, classified named entities and temporal expressions. The authors in Bamman and Smith (2014) also extract 927,403 biographies in this case from the English Wikipedia. The set is pre-processed in order to learn event classes in biographies.

Regarding the automatic extraction of parallel corpora, Wikipedia has been traditionally used as a resource. In Adafre and de Rijke (2006), the authors extract parallel sentences based on the available metadata in Wikipedia texts. Both Yasuda and Sumita (2008) and Plamada and Volk (2012) extracted parallel sentences by translating the articles into a common language and consider those sentences with a high translation quality to be parallel. The ACCURAT project (Ştefănescu et al., 2012; Skadiņa et al., 2012) also devoted efforts in parallel sentence mining in Wikipedia. Later, Barrón-Cedeño et al. (2015) used the combination of cross-lingual similarity measures to extract domain specific parallel sentences. The most recent initiative is the so-called LASER (Artetxe and Schwenk, 2019b), which relies on vector representations of sentences to extract similar pairs. This toolkit has been used to extract the WikiMatrix corpus (Schwenk et al., 2019) which contains 135 million parallel sentences for 1,620 different language pairs in 85 different languages.

As far as we are concerned, there is no gender-balanced dataset for machine translation, except for the artificial gold standard created for English–Spanish (Font and Costa-jussà, 2019). However, there are a notable number of works towards doing research in the area: from balancing data sets in monolingual tasks (Webster et al., 2018; Rudinger et al., 2018) to evaluating gender bias in several tasks (Basta et al., 2019; Stanovsky et al., 2019).

3. GeBioToolkit

3.1. Base Architecture

To extract the corpus previously introduced, we develop GeBioToolkit. The tool retrieves (multi-)parallel sentences for any Wikipedia category and for any arbitrary number of languages at the same time. In its default configuration, the toolkit retrieves gender-balanced corpora. GeBioToolkit is composed by three blocks. The first block, corpus extractor, provides a layer to transform, collect and select entries in the desired languages. The second block, corpus alignment, finds the parallel sentences within a text and provides a quality check of the parallel sentences given a few restrictions. The third block, gender classifier, includes a filtering module which classifies the gender of the entry and outputs the final parallel corpus. The gender detection functionality can be activated or deactivated at will, allowing the tool to be used in a more general context, where gender-balanced data is either non-relevant or not needed. Figure 1 depicts the architecture.

The tool requires three inputs: (*i*) a list of the desired languages, (*ii*) the dump files for the languages[4] and (*iii*) a list of the articles' titles belonging to the category to extract (currently in English). For our purpose of gathering a corpus of biographies, we retrieve the list of articles that belong to the "living people" category by using the PetScan tool[5].

Given an input article, the **corpus extractor** module starts by looking for equivalent articles in the other languages via the Wikipedia inter-language links. The module also retrieves the multilingual titles providing a dictionary of

[2] https://wit3.fbk.eu/

[3] https://www.h-its.org/en/research/nlp/wikibiography-corpus/

[4] Dumps can be downloaded from https://dumps.wikimedia.org/

[5] https://petscan.wmflabs.org/

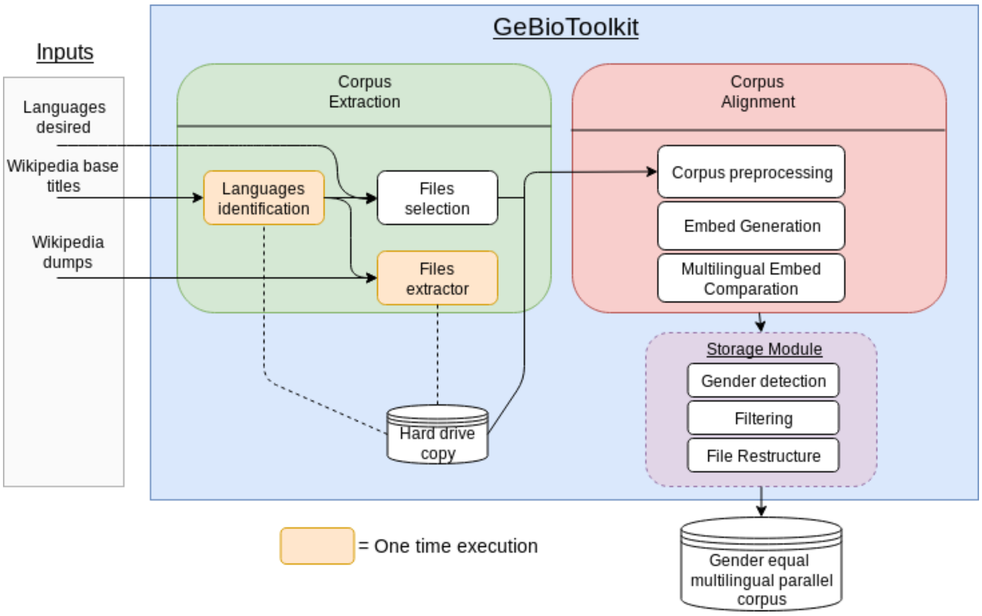

Figure 1: GeBioToolkit architecture

{english_entry_title: [language, title]} that is used on the file extraction and file selection modules. File extraction retrieves then the selected entries in the previous dictionary from the Wikipedia dump. GeBioToolkit uses a modified version of the *wikiextractor*[6] software to retrieve and store the different Wikipedia entries from each language. Finally, file selection generates a dictionary similar to the one obtained before, but it only stores the entries for which the files were successfully retrieved.

The **corpus alignment** module makes use of the text and dictionaries retrieved in the previous step together with the LASER toolkit[7]. LASER (Language-Agnostic SEntence Representations) allows to obtain sentence embeddings through a multilingual sentence encoder (Schwenk et al., 2019). Translations can be found then as close pairs (tuples) in the multilingual semantic space. In order to perform the parallel sentence extraction, we follow the margin-based criterion introduced in Artetxe and Schwenk (2019a). The margin between two candidate sentences x and y is defined as the ratio between the cosine distance between the two embedded sentences, and the average cosine similarity of its nearest neighbors in both directions:

$$margin(x,y) = \frac{\cos(x,y)}{\sum\limits_{z \in NN_k(x)} \frac{\cos(x,z)}{2k} + \sum\limits_{z \in NN_k(y)} \frac{\cos(y,z)}{2k}},$$

where $NN_k(x)$ denotes the k unique nearest neighbors of x in the other language, $NN_k(y)$ the same for y.

To extract parallel sentences on more than two languages, let us say i languages, we use a greedy approach with a pivot language L_1. We detect all the parallel sentences in the pairs L_1–L_i and then extract the intersection of sentences between the language pairs.

3.2. Gender Detection Module

The previous two blocks implement a general approach to parallel sentence extraction following a similar methodology as that used to extract the WikiMatrix corpus. But for our purpose, we need to specifically deal with gender bias. Wikipedia is known to have a gender bias in its content. Depending on the language and year of the study, the percentage of biographies of women with respect to the total of biographies has been reported to range from a 13% to a 23% (Reagle and Rhue, 2011; Bamman and Smith, 2014; Wagner et al., 2016). And it is not only this, but also men and women are characterised differently (Graells-Garrido et al., 2015; Wagner et al., 2016), showing for instance more man in sports and woman in divorces (Reagle and Rhue, 2011).

To allow extracting a gender-balanced dataset, we detect the gender automatically and filter files in order to have 50% of articles for each gender. Following Reagle and Rhue (2011), the gender of the article is extracted as that corresponding to the maximum number of gendered pronouns (i.e., he and she in English) mentioned in the article. According to their results, this method of gender inference is overwhelmingly accurate, in a random test set of 500 articles, it achieved 100% precision with 97.6% recall (12 articles had no pronominal mentions and so gender was not assigned).

3.3. Cleaning Module

We analyse the accuracy of the extractions in Section 4.3., but a simple visual inspection already shows a frequent source of noise in the extracted data that can be easily removed with a post-processing step. Some of the extracted sentence pairs include information in one language that is lacking in the other one(s). For example, the sentence in Spanish *”Mahrez se casó con su novia inglesa en 2015 y tuvieron una hija ese mismo año.”* is aligned with the sentence in English *”Mahrez married his English girlfriend*

[6] https://github.com/attardi/wikiextractor

[7] https://github.com/facebookresearch/LASER

```
<doc docid="Aurelia Arkotxa " wpid="51690640" language="en" topic="C6" gender="Female" >
<title>Aurelia Arkotxa </title>
<seg id="1">She teaches classics at the University of Bayonne; she was co-founder of the literary magazine and a new newspaper.<\seg>
</doc>
<doc docid="Catriona Gray " wpid="51838666" language="en" topic="C2" gender="Female">
<title>Catriona Gray </title>
<seg id="1">In addition, she obtained a certificate in outdoor recreation and a black belt in Choi Kwang-Do martial arts.<\seg >
<seg id="2>Catriona Elisa Magnayon Gray (born 6 January 1994) is a Filipino-Australian model, singer, and beauty pageant titleholder who was crowned Miss Universe 2018.<\seg>
<seg id="3">Gray was born in Cairns, Queensland, to a Scottish-born father, Ian Gray, from Fraserburgh, and a Filipina mother, Normita Ragas Magnayon, from Albay.<\seg >
</doc>
```

```
<doc docid="Aurelia Arkotxa" wpid="7789214" language="es" topic="C6" gender="Female" >
<title>Aurelia Arkotxa</title >
<seg id="1">Enseña cultura clásica en la facultad de Bayona; fue cofundadora de una revista literaria y de un diario.<\seg>
</doc>
<doc docid="Catriona Gray" wpid="8411924" language="es" topic="C2" gender="Female" >
<title>Catriona Gray </title >
<seg id="1">Además, obtuvo un Certificado en Recreación al Aire Libre y un cinturón negro en Artes Marciales de Choi Kwang-Do.<\seg>
<seg id="2">Catriona Elisa Magnayon Gray (6 de enero de 1994) es una modelo y reina de belleza australiana-filipina, ganadora de Miss Universo 2018 representando a Filipinas.<\seg>
<seg id="3">Gray nació en Cairns, Queensland de un padre australiano nacido en Escocia, Ian Gray, de y una madre filipina, Normita Ragas Magnayon, de Albay.<\seg>
</doc>
```

```
<doc docid="Aurelia Arkotxa" wpid="1473820" language="ca" topic="C6" gender="Female" >
<title>Aurelia Arkotxa</title >
<seg id="1">Ensenya cultura clàssica a la Universitat de Baiona; va ser cofundadora d'una revista literària i d'un diari.<\seg>
</doc>
<doc docid="Catriona Gray" wpid="1635999" language="ca" topic="C2" gender="Female" >
<title>Catriona Gray </title >
<seg id="1">També va obtenir un Certificat en Recreació a l'aire lliure i un cinturó negre en Arts Marcials de Choi Kwang-Do.<\seg>
<seg id="2">Catriona Elisa Magnayon Gray (6 de gener de 1994) és una model i reina de bellesa filipina d'origen australià, guanyadora de Miss Univers 2018.<\seg>
<seg id="3">Gray va néixer a Cairns, Queensland d'un pare australià nascut a Escòcia, Ian Gray, de , i d'una mare filipina, Normita Ragas Magnayon, de Albay.<\seg>
</doc>
```

Table 1: Example of three documents extracted in English (top) and the parallel counterparts in Spanish (middle) and Catalan (bottom) from GeBioCorpus-v2.

Rita Johal in 2015.", where the Spanish segment *"y tuvieron una hija ese mismo año."* is not present in English. To avoid this, we filter out sentence pairs with large length ratio (filter samples in which one of the sentences is at least 20% longer than the others). Such filtering can be generalised to distant language pairs by estimating language-dependent length factors (Pouliquen et al., 2003) or considering more elaborated corpus filtering techniques (Koehn et al., 2019).

3.4. File Restructure Module

Finally, the extracted parallel sentences are written as an xml file using a document-level mark-up. As said before, output documents do not preserve coherence any more, but document-level characteristics such as lexical consistency are kept and the mark-up allows use the information.

Table 1 shows an excerpt of the output file for English and Spanish. For each original biography, we keep the English title as document ID, the original Wikipedia ID, the extracted gender, and in our clean test set (see Section 4.) we also add the topic or category of each document. The `docid` field links documents among languages.

4. GeBioCorpus

We use the GeBioToolkit presented in the above section to extract a multilingual parallel corpus balanced in gender for English, Spanish and Catalan. The motivation for extracting this specific dataset is the variation in morphology in English (Germanic language) as compared to Spanish and Catalan (Romance languages). The variation in morphology is one of the most relevant challenges when solving biases in gender, see examples in (Prates et al., 2019; Font and Costa-jussà, 2019).

4.1. Data Statistics

We extract all the articles belonging to the category "living people" in the 40 largest Wikipedia editions to have a rough idea of the number of biographies available per language. Figure 2 shows the amount of these articles for the 20 languages with the largest number of biographies. The edition with the most available entries is the English one with 922,120 entries. The Spanish Wikipedia contains 148,445 entries (6th largest one) and the Catalan edition 40,983 (20th largest one). All three are highlighted in Figure 2. Even if the Catalan Wikipedia is not as big as the English and Spanish ones, there is a noticeable amount of comparable articles between Spanish and Catalan which translates into significant number of parallel sentences —Schwenk et al. (2019) extracted in WikiMatrix 3,377 k sentences for en–es, 1,580 k sentences for ca–es and 210 k sentences for en–ca from the full editions.

GeBioToolkit extracts multilingually aligned parallel sentences, so it is interesting to study also the union of languages. The more languages involved, the lesser number or comparable articles one will obtain. Figure 3 shows the number of articles per sets of languages and broken-down by gender. The total number of documents decays when starting with English and Arabic (set with only 2 languages) and one incrementally adds French, German, Italian, Spanish, Polish, Russian, Portuguese, Japanese and Dutch. With all languages (set with 11 languages), the number of comparable documents results in 17,285, out of which 13,676 are male bios and 3,609 are female bios. These numbers reconfirm the results by (Reagle and Rhue, 2011; Bamman and Smith, 2014; Wagner et al., 2016) on the underrepresentation of women in Wikipedia. Note that by doing the union of different languages and the intersection of its ar-

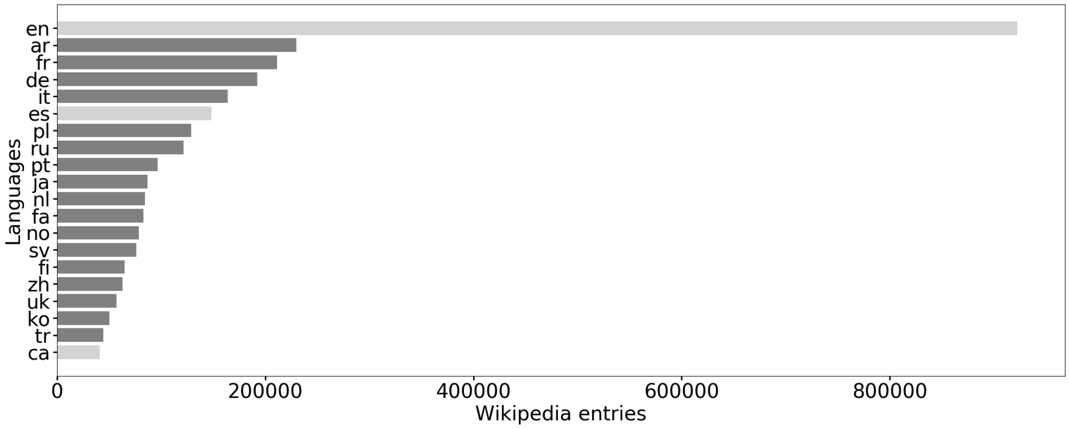

Figure 2: Number of Wikipedia entries under the "living people" category for the 20 Wikipedia editions with the most number of entries. Light gray marks the languages used in our study.

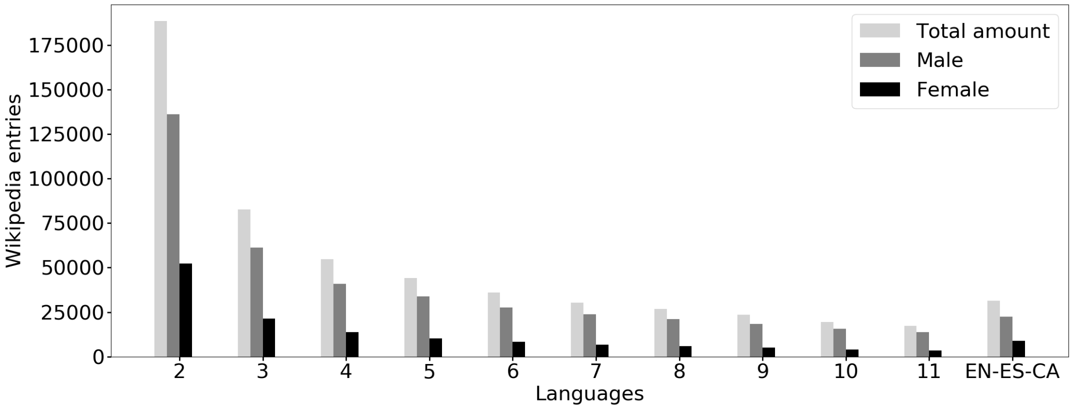

Figure 3: Number of documents by sets of languages under the Wikipedia category "living people" distributed by gender. See Section 4.1. for the specific languages involved in the sets.

EN	In 2008, she was honored by the Angolan Ministry of Culture with a certificate of honor for her services to Angolan children's literature.
ES	En 2008, fue honrada por el Ministerio de Cultura de Angola con un certificado de honor por sus servicios a la literatura infantil angoleña.
CA	El 2008 va rebre el premi del Ministeri de Cultura angolès amb un certificat d'honor pels seus serveis a la literatura infantil angolesa.
EN	His poetry is characterized by a colloquial language and by his reflections regarding every day events or situations.
ES	Su poesía se caracteriza por un lenguaje coloquial y por la reflexión a partir de acontecimientos o situaciones cotidianas.
CA	La seva poesia es caracteritza per un llenguatge col·loquial i per la reflexió a partir d'esdeveniments o situacions quotidianes.
EN	Bridegroom was an actor and songwriter who hosted the TV series "The X Effect".
ES	Bridegroom era un actor y compositor quién participó de la serie de televisión "El Efecto X.
CA	Bridegroom era un actor i compositor que va participar de la sèrie de televisió "The X Effect."
EN	She was also recognised in 2015 by the British Council for her community work, and by the BBC as part of their "100 Women" series.
ES	También fue reconocida en 2015 por el British Council por su trabajo comunitario, y por la BBC como parte de su serie "100 Women (BBC)".
CA	També va ser reconeguda el 2015 pel British Council pel seu treball comunitari, i per la BBC com a part de la seva sèrie "100 Women (BBC)".

Table 2: Examples of parallel segments extracted for English, Spanish and Catalan from GeBioCorpus-v2.

ticles, the amount of documents decays quite abruptly, but the percentual difference between man and woman remains close to be constant.

When we consider English, Spanish and Catalan, we retrieve 31,413 bibliographies. GeBioToolkit selects 81,405 sentences from them, 53,389 sentences are related to male bibliographies and 28,016 to female bibliographies. 47.5% of the male sentences are removed to obtain a gender-balanced corpus. Then, GeBioToolkit performs the length-based cleaning, which filters the corpus down to 16,679 sentences on male bibliographies and 10,730 sentences on female bibliographies.

As output, we provide two versions of the corpus. GeBioCorpus-v1 contains 16,000 sentences as ex-

GeBioCorpus-v1	En		Es		Ca	
	F	M	F	M	F	M
Documents	2287	2741	2287	2741	2287	2741
Sentences	8000	8000	8000	8000	8000	8000
Average sent/doc	3.5	2.9	3.5	2.9	3.5	2.9
Words	228.9k	235.3k	230.1k	236.0k	240.9k	245.7k
Average words/doc	56.9	51.1	56.3	47.3	59.8	53.3
Vocabulary	24.3k	24.8k	27.1k	27.6k	27.6k	28.0k

GeBioCorpus-v2	En		Es		Ca	
	F	M	F	M	F	M
Documents	257	339	257	339	257	339
Sentences	1000	1000	1000	1000	1000	1000
Average sent/doc	3.9	2.9	3.9	2.9	3.9	2.9
Words	28.7k	27.5k	28.9k	28.0k	30.1k	29.4k
Average words/doc	61.6	46.6	61.4	46.4	65.4	49.7
Vocabulary	6.1k	6.0k	6.5k	6.5k	6.7k	6.6k

Table 3: Statistics for the two corpora extracted by GeBioToolkit without (GeBioCorpus-v1) and with (GeBioCorpus-v2) manual post-edition on the raw output. Figures are given per languages (En, Es, Ca) and gender (F, M).

tracted from Wikipedia without any manual post-edition. GeBioCorpus-v2 contains 2,000 sentences with manual post-edition as explained in Section 4.2.. One set excludes the other in addition to excluding other sentences (1,730 for female) that we plan to post-edit in the future. See statistics on number of documents, sentences, words and vocabulary broken-down per language and gender in Table 3. Table 2 provides some examples of sentences from GeBioCorpus-v2.

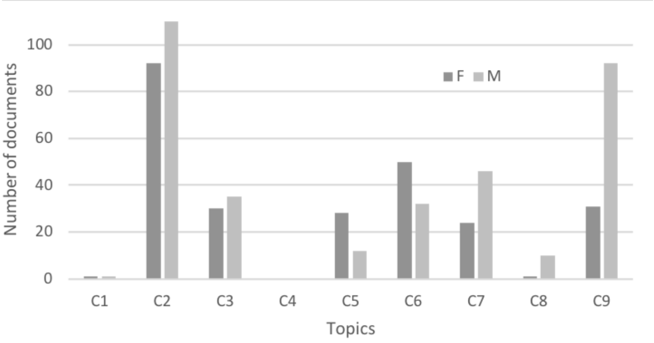

Figure 4: Distribution of topic/categories for the manually annotated test set. The distribution is further split by gender (articles of women (F) and men (M)). See Section 4.2. for the categories defined.

4.2. Manual Post-Editing and Categorization

After the extraction of the complete corpus, and in order to provide a high-quality evaluation dataset, we ask native/fluent speakers to post-edit 2,000 sentences belonging to both female and male documents. Annotators were asked to either edit segments to make them parallel (if differences were small) or to delete segments that were too different. Also during the process, annotators were asked to provide a topic for the profession of each personality. Categories are based on the Wikipedia list of occupations[8], which includes: (C1) Healthcare and medicine, (C2) Arts, (C3) Business, (C4) Industrial and manufacturing, (C5) Law enforcement, social movements and armed forces, (C6) Science, technology and education, (C7) Politics, (C8) Religion, and (C9) Sports. This corresponds to the field ”topic” added in the header of each document as seen in Table 1. Figure 4 shows the distribution of the 9 topics by gender. Note that the most unbalanced topics are C5 (Law enforcement, social movements and armed forces), C8 (Religion) and C9 (Sports). We report other statistics for this dataset in Table 3.

4.3. Human Evaluation

Finally, we perform a human evaluation on the quality of GeBioCorpus-v1, that is, the data as extracted by GeBioToolkit without any post-edition. For this, we randomly select 50 sentences in the three languages (English, Spanish and Catalan) and present each tuple to 7 different native/fluent speakers. These annotators were asked to score a tuple with 1 if the three sentences convey the same meaning and 0 otherwise. On average, evaluators gave 87.5% accuracy and when computing the majority vote among the evaluators accuracy reached 96%. We computed Fleiss' kappa (Fleiss and others, 1971) as a measure for the interannotator agreement which resulted in 0.67, which is considered a substantial agreement (Landis and Koch, 1977).

5. Conclusions

This paper presents two main contributions. On one side, GeBioToolkit, based on LASER, allows to extract automatically multilingual parallel corpora at sentence-level and can be customized in number of languages and balanced in gender. Document-level information is kept in the cor-

[8] https://en.wikipedia.org/wiki/Lists_of_occupations

pus and each document is tagged with the ID of the original Wikipedia article, the language, and the gender of the person it is referring to. On the other side, we provide a multilingual corpus of biographies in English, Spanish and Catalan, GeBioCorpus. Two versions of this corpus are presented. The first version, GeBioCorpus-v1, contains 16k sentences which have been directly extracted using GeBioToolkit. This version of the corpus is used to make an evaluation of the quality of the extractions produced by our tool. A manual evaluation shows that the accuracy in the multilingual parallel sentences is 87.5%. The second version of this corpus, GeBioCorpus-v2, contains 2k sentences which have been post-edited by native speakers and categorized with type of profession. In this case, and in addition to the automatic tags of each document (ID, language and gender), each document is tagged with an occupation category. This labels allows to split the corpus in different subsets to evaluate domain-specific translations and gender accuracy for instance.

As future improvements to the tool, we will remove the dependence on PetScan —or any external tool to obtain lists of articles—, and extract all the necessary information from the input Wikipedia dump. We will also study the viability of using Wikidata information instead of the most frequent pronoun in a text in order to classify the gender of an article. Finally, we want to perform a detailed error analysis for GeBioCorpus-v1 to investigate which module from the toolkit causes errors, and and analysis of the quantity and content of post-edition in GeBioCorpus-v2 to estimate the cost of post-edition.

Both GeBioToolkit and GeBioCorpus are available in github[9].

Acknowledgments

The authors want to thank Jordi Armengol, Magdalena Biesialska, Casimiro Carrino, Noe Casas, Guillem Cortès, Carlos Escolano, Gerard Gallego and Bardia Rafieian for their help in postediting the GeBioCorpus, Eduardo Lipperheide, Anna Li, Ayyoub Marroun, Laura Termes, Adrià Garcia, Alba Salomon and Clara Grenzner for the manual evaluation and Melissa Steiner, Ekaterina Lapshinova-Koltunski and Pauline Krielke for all the support and fruitful discussions.

This work is supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund through the postdoctoral senior grant Ramón y Cajal and by the Agencia Estatal de Investigación through the projects EUR2019-103819 and PCIN-2017-079. It is also partially funded by the German Federal Ministry of Education and Research under the funding code 01IW17001 (Deeplee). Responsibility for the content of this publication is with the authors.

6. Bibliographical References

Adafre, S. 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.

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019b). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. volume 7, pages 597–610. MIT Press, September.

Bamman, D. and Smith, N. A. (2014). Unsupervised Discovery of Biographical Structure from Text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Barrón-Cedeño, A., España-Bonet, C., Boldoba, J., and Màrquez, L. (2015). A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 3–13, Beijing, China, July.

Basta, C., Costa-jussà, M. R., and Casas, N. (2019). Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, August. Association for Computational Linguistics.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16^{th} Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Costa-jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1.

Ştefănescu, D., Ion, R., and Hunsicker, S. (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 .

Fleiss, J. et al. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Font, J. E. and Costa-jussà, M. R. (2019). Equalizing gender biases in neural machine translation with word embeddings techniques. *CoRR*, abs/1901.03116.

Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First Women, Second Sex: Gender Bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext &#38; Social Media*, HT '15, pages 165–174, New York, NY, USA. ACM.

Habash, N., Zalmout, N., Taji, D., Hoang, H., and Alzate, M. (2017). A Parallel Corpus for Evaluating Machine Translation between Arabic and European Languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 235–241, Valencia, Spain, April. Association for Computational Linguistics.

Koehn, P., Birch, A., and Steinberger, R. (2009). 462 Machine Translation Systems for Europe. In *Proceedings*

[9] https://github.com/PLXIV/Gebiotoolkit

of the Twelfth Machine Translation Summit, pages 65–72. Association for Machine Translation in the Americas, AMTA.

Koehn, P., Guzmán, F., Chaudhary, V., and Pino, J. (2019). Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy, August. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. In Ellen Riloff, et al., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4791–4796. Association for Computational Linguistics.

Park, J. H., Shin, J., and Fung, P. (2018). Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, October-November. Association for Computational Linguistics.

Plamada, M. and Volk, M. (2012). Towards a Wikipedia-extracted alpine corpus. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora: Language Resources for Machine Translation in Less-Resourced Languages and Domains*, pages 81–87, Istanbul, Turkey, May.

Pouliquen, B., Steinberger, R., and Ignat, C. (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.

Prates, M., Avelar, P., and Lamb, L. (2019). Assessing Gender Bias in Machine Translation – A Case Study with Google Translate. *Neural Computing and Applications*, 03.

Reagle, J. and Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communication*, 5(0).

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender Bias in Coreference Resolution. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *CoRR*, abs/1907.05791.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiș, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., and Pinnis, M. (2012). Collecting and using comparable corpora for statistical machine translation. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1):5, Mar.

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *ArXiv e-prints*, October.

Yasuda, K. and Sumita, E. (2008). Method for Building Sentence-Aligned Corpus from Wikipedia. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 64–66, Menlo Park, CA.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.