Creating Clustered Comparable Corpora from Wikipedia with Different Fuzziness Levels and Language Representativity

Anna Laskina, Eric Gaussier, Gaelle Calvary

Creating Clustered Comparable Corpora from Wikipedia with Different Fuzziness Levels and Language Representativity

Anna Laskina, Eric Gaussier, Gaelle Calvary
Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG
38000 Grenoble, France
{anna.laskina, eric.gaussier, gaelle.calvary}@univ-grenoble-alpes.fr

Abstract

This paper is dedicated to the extraction of clustered comparable corpora from Wikipedia, that is comparable corpora with labelled information corresponding to the topics associated to each document. Despite the importance of such corpora for evaluating text clustering and classification methods in the context of comparable corpora, there is a notable absence of automatic algorithms capable of creating them with adjustable fuzziness levels and language representativity. The methodology we propose here offers control over the cluster distribution across languages, enables fine-tuning of fuzziness levels, and facilitates customization to accommodate specific subject areas. Moreover, we have developed a dedicated tool specifically designed for our purpose and present 18 bilingual clustered comparable corpora spanning English, French, German, Russian, and Swedish languages. The analysis of these corpora demonstrates the effectiveness and flexibility of the approach in constructing corpora with varying levels of fuzziness and language representativity. Our results, tool and corpora, pave the way to construct various gold standard collections for future research in clustering and classification in comparable corpora.

Keywords: Comparable corpora, Gold standard colelctions, Text clustering/text classification, Wikipedia

1. Introduction

Our objective in this study is to provide a tool to automatically extract from Wikipedia comparable corpora with clustering information, each cluster corresponding to a meaningful Wikipedia category called *topic* afterwards. We refer in the remainder to such corpora as *clustered comparable corpora*. As such, this study is part of a broader initiative on clustering comparable corpora where gold standard collections are needed in order to compare different clustering approaches and methods. It has to be noted that such collections can also be used for studying, in a (semi-)supervised setting, text classification in comparable corpora.

Wikipedia stands out as a well-known and freely available public resource, offering a vast array of texts in multiple languages. Moreover, the texts in Wikipedia covering similar topics are intricately linked and categorized in the same high-level categories, facilitating the construction of coherent and comprehensive comparable corpora. In addition, as many articles cover different topics and belong to different Wikipedia categories, it is possible to construct clustered comparable corpora in which documents can have different *fuzziness levels*, *i.e.*, be assigned to one or more clusters, enabling more nuanced analysis and interpretation of the data. Lastly, in, for example, the context of bilingual comparable corpora involving two languages ℓ_1 and ℓ_2, for a given set of topics, it is possible to extract from Wikipedia different clustered comparable corpora with different proportions of clusters containing only documents in ℓ_1, only documents in ℓ_2, or a mixture of documents in ℓ_1 and documents in ℓ_2.

Based on the above considerations, we aim in this study at developing a methodology and an associated suite of tools to extract clustered comparable corpora from Wikipedia while offering to researchers the possibility to:

- Tailor such corpora to specific subjects,
- Regulate their fuzziness levels,
- Control the proportions of monolingual and multilingual clusters.

Through the integration of these elements, researchers can access richer, more diverse datasets, thereby advancing the frontiers of data-driven inquiry and analysis in comparable corpora. In particular, they can use the collected datasets for evaluating clustering and/or classification methods. The described methodology can be applied to any other knowledge base with a similar structure to Wikipedia when the need arises to create collections with different knowledge from Wikipedia. For simplicity, we focus here on the construction of clustered *bilingual* comparable corpora. The extension to multilingual corpora is nevertheless direct.

The structure of this paper is organized as follows. Section 2 provides an overview of related work in the field of extracting comparable corpora from Wikipedia. Section 3 presents our proposed methodology, which consists of two main components: extrcating a category tree from the Wikipedia cateogry graph (Section 3.1) and building clustered bilingual comparable corpora (Section 3.2). Section 4 presents our results, consisting of a tool we

BUCC 2024: The 17th Workshop on Building and Using Comparable Corpora, pages 85–93
20 May, 2024.

developed (Section 4.1) and an analysis of several collected corpora (Section 4.2). Finally, Section 5 concludes the paper by summarizing the key findings and outlining potential avenues for future research in this area.

2. Related Work

Wikipedia is widely used across different domains, making it a suitable primary data source for extracting comparable corpora. Several studies have utilised Wikipedia data for dictionary extraction (Chu et al., 2014; Erdmann et al., 2008; Yu and Tsujii, 2009) and machine translation tasks (Ramesh and Sankaranarayanan, 2018; Ruiter et al., 2019; Alegria et al., 2013). Wikipedia data is commonly used to train pre-trained models, in particular word embeddings and language models. Examples of pre-trained word representations that use Wikipedia text corpora include fastText (Mikolov et al., 2018), BERT (Devlin et al., 2019) and LASER (Artetxe and Schwenk, 2019). Recent advancements in large language models (LLMs), such as GPT-3 (Brown et al., 2020), mT6 (Chi et al., 2021), llama (Touvron et al., 2023), and LaMDA (Thoppilan et al., 2022), have been also incorporating Wikipedia data into their training processes.

There are several works in Wikipedia-based comparable corpora. Although some efforts concentrate on collecting parallel sentences (Plamada and Volk, 2012; Plamadă and Volk, 2013) or pairs of articles (Saad et al., 2013; Goyal et al., 2020) in multiple languages to create comparable corpora, these endeavors are mainly aimed at machine translation applications rather than clustering and classification tasks, which are aligned with our objectives.

When exploring methodologies for creating comparable collections from Wikipedia, various works strive to gather comparable collections for a specific language pair and a specific topic. These works vary mainly in their document selection process for the chosen topic. In (Otero and López, 2010; Otero et al., 2011), the authors align topics with specific Wikipedia categories and considered three possible options for the comparability of documents in different languages: documents belonging to the same topic because they have the same associated category (non-aligned), documents connected by an inter-language link (softly-aligned), and documents connected by an inter-language link and belonging to the same category (strongly-aligned). A limitation of this approach is that it focuses on documents directly related to the selected category, which limits the size of the corpus and poses challenges in assembling larger corpora. According to (Barrón-Cedeno et al., 2015), an alternative approach involves selecting documents that are not only directly related to a topic associated with Wikipedia categories but also those associated with its subcategories. We intend to adopt this strategy. A recent study (España-Bonet et al., 2023) proposed an approach to improve the selection of documents from subcategories of the Wikipedia category associated with the topic. This was achieved by using a vocabulary that describes the topic and retaining only those subcategories whose titles contain at least one word from the vocabulary. While acknowledging its advantages, we have decided not to employ this approach in this paper. This is mainly caused by the topic vocabularies, which can number over a hundred and vary depending on the collection topic, fuzziness levels, and language representation. Nevertheless, we do intend to explore its potential inclusion in future work. That said, none of these methods aims at building clustered comparable corpora and the methodology we propose in this paper is the first one, as far as we know, to address this problem.

3. Methodology

We describe in this section the methodology followed to extract clustered bilingual comparable corpora from Wikipedia. It relies on a first step that creates a category tree from the Wikipedia category graph to determine appropriate topics for labeling a corpus. The second step involves creating the corpus according to the specified preferences regarding language representativity and fuzziness.

3.1. From a Category Graph to a Category Tree

Each page in Wikipedia typically has multiple categories, which are organised into a hierarchical graph known as the Wikipedia category graph (Hecht and Gergle, 2010). The Wikipedia category graph, which has been the subject of many studies (Zesch and Gurevych, 2007; Suchecki et al., 2012; Aspert et al., 2019), contains numerous cycles (España-Bonet et al., 2023; Barrón-Cedeno et al., 2015) in the sense that a category can refer to itself as a parent category after several generations. For instance, the category *Soil* serves as both a parent and a subcategory of the category *Soil science*, creating a cycle of *Soil* $\rightarrow$ *Soil science* $\rightarrow$ *Soil*. This said, it has been shown that the Wikipedia category graph can be hierarchized by identifying a root and organising the graph into hierarchy levels according to the length of directed paths from the root (Aouicha et al., 2016; Aghaebrahimian et al., 2022). Following this idea, the category *Main Topic Classifications* has often been chosen as the root category and has therefore been assigned a level of 0. Note that this category was selected because it includes the main Wikipedia

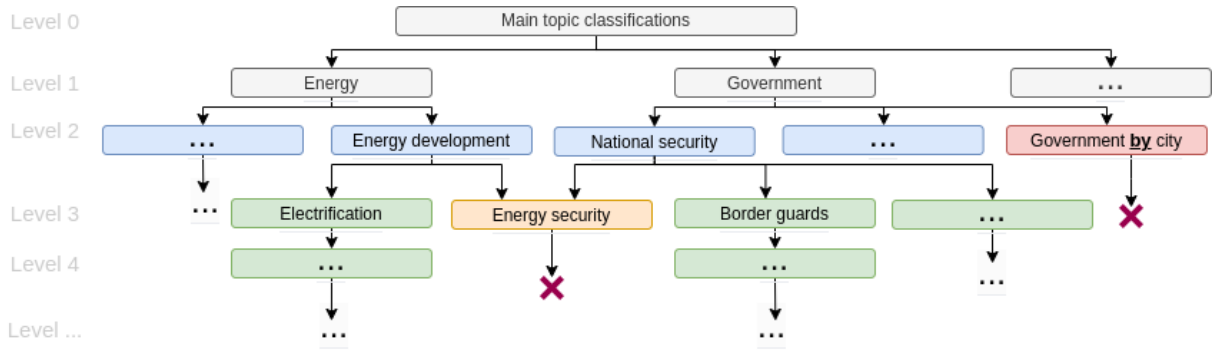

Figure 1: Part of the category tree showing the relationships between categories and their types: *insignificant* (red), *ambiguous* (orange) and *theme* (blue, green and gray). When the level of reference equals 2, the blue *theme* categories generate clusters in the corpus, while the green *theme* categories are used when searching for clusters in documents. The grey *theme* categories are not used in corpus generation because they are too general.

topics for categorisation, as provided by Wikipedia itself.

Several studies have further demonstrated that not all categories are adequate for creating a cluster (Aghaebrahimian et al., 2022). Relevant categories can either be selected manually from a list, typically limited to a few dozen categories (Plamadă and Volk, 2013; Barrón-Cedeno et al., 2015), or extracted automatically by filtering out non relevant categories. The latter option is the most promising (España-Bonet et al., 2023) and is the one we adopt here.

The first aspect which differentiates relevant and non relevant categories relates to the fact that some categories indeed represent specific topics (*e.g.*, *Music*, *History*), while others mainly serve for organizing the whole Wikipedia collection (*e.g.*, *Outlines of general reference*). The latter categories, precisely defined below, are not suitable for clustering documents into topics and are referred to here as *insignificant* categories.

Definition 3.1 *A category, the name of which contains any of the words by, in, from, about, and, after, list, award, image, quotation, event, outline, redirect, people is called an* insignificant *category.*

The second aspect relates to the fact that the Wikipedia category graph contains duplicates of categories with different parents at the same distance from the root. The distance considered here is the length of the shortest path from a given node to the root. These categories are *ambiguous* in that they are equally relevant to all of their parent categories equally distant from the root and do not exhibit a stronger association with any one of them.

Definition 3.2 *A category that has more than one parent category equally distant from the root category is called* ambiguous.

We focus here on all categories but insignificant and ambiguous categories for creating clustered comparable corpora. Such categories are called *theme* categories in the remainder of the paper:

Definition 3.3 *A category witch is neither* insignificant*, nor* ambiguous *is called a* theme *category.*

In Figure 1, the category *Government by city* is *insignificant* because it contains the word *by*, the category *Energy security* is *ambiguous* as it is both a subcategory of *Energy development* and *National security*, which are both level 2 categories, and the categories *Electrification* and *National security* are *theme* categories because they are neither *insignificant* nor *ambiguous*.

Finally, the process we rely on to create a category tree $\mathcal{T}$ aims at filtering out the Wikipedia category graph by removing cycles and relying only on theme categories, thus obtaining a tree backbone of the Wikipedia category graph fully suited to topical clustering/classification. It goes as follows:

1. Set the root node c_0 of $\mathcal{T}$ to the category *Main Topic Classifications*; set the level l to 0.
2. Recursively add to $\mathcal{T}$, in a breadth-first manner and at level $(l+1)$, all subcategories of all theme categories present at level l in $\mathcal{T}$ if they are not already in $\mathcal{T}$; mark as insignificant and ambiguous the added subcategories complying with definitions 3.1 and 3.2.

Note that one can easily check if a category is ambiguous by verifying if it is already present in $\mathcal{T}$ at the same level. This process naturally stops at level 22.

Table 1 displays the different types of categories at the different levels of the tree obtained by the above process. As one can note, there are 39 theme categories at level 1, 825 at level 2 and 5539 at level 3. To ensure homogeneity between clusters in the final corpus, we consider as original

Level	*insignificant*	*ambiguous*	*theme*
1	2	0	39
2	341	46	825
3	1542	921	5539
4	4814	2690	16914
5	12032	4769	38390
6	25251	12647	54742
7	29710	8896	69122
8	35695	10195	59671
9	28389	6392	53236
10	23272	5759	41600
11	19065	2527	27797
12	9767	1039	15472
13	4145	415	10345
14	4050	317	6440
15	1852	177	2541
16	1172	78	1275
17	342	19	332
18	83	9	54
19	4	0	9
20	0	0	3
21	0	0	4
22	0	0	0

Table 1: The amount of *insignificant*, *ambiguous* and *theme* categories in the category tree by level. From level 22 there are no more *theme* categories.

topics to construct clustered bilingual comparable corpora theme categories at the same level, which will be referred to as l_r for *level of reference*:

Definition 3.4 *A* theme *category at level l_r in the category tree is called a* topic*. Furthermore, any Wikipedia category c as well as the Wikipedia documents assigned to it belong to a topic t if $c = t$ or c is a descendant of t in $\mathcal{T}$.*

Different levels of reference can be used depending on the balance between coarse-grained and fine-grained clusters one is interested in. Lastly, the clusters (or classes if one is rather interested in text classification) we consider for constructing clustered (or categorized) bilingual comparable corpora are a subset of the topics defined above. As described below, we will rely on both primary and secondary topics to obtain our clusters.

3.2. Corpus Creation

We consider here the creation of a clustered bilingual comparable corpus where documents are written in either language ℓ_1 or language ℓ_2. Such a corpus can display three types of clusters: clusters of type 1 (resp. 2) containing only documents written in ℓ_1 (resp. ℓ_2) and documents of type $1\&2$ containing both documents written in ℓ_1 and documents written in ℓ_2. Of course, all types may not be represented in every clustered bilingual comparable corpus; in addition, for simplicity, we focus here on clustered corpora in which a document can only belong to clusters of the same type.

In order to control the representativity of each language in the corpus to be created, we define three hyperparamters, denoted Nt_1, Nt_2 and $Nt_{1\&2}$, which specify the number of *primary* topics of type 1, 2 and $1\&2$ one is interested in. Each number Nt_1, Nt_2 and $Nt_{1\&2}$ can either be set directly or be randomly chosen from a set of values defined by the user (see Table 2 for example). Furthermore, we allow users the possibility to have clusters of different sizes by randomly selecting, from a given set of values, the number of documents Nd_{ij} associated to the j^{th} topic of type i ($i \in \{1, 2, 1\&2\}$).

In addition to controlling the language representativity, we also want to control the overall degree of fuzziness of documents across clusters. To this end, we introduce two additional hyperparameters, f_{min} and f_{max}, which respectively represent the minimum and maximum numbers of clusters a document should belong to. f_{min} is lower bounded by 1 and upper bounded by f_{max}, whereas f_{max} is lower bounded by f_{min} and upper bounded by the maximum number of topics a document can have in Wikipedia. Both f_{min} and f_{max} are defined by the user.

Collecting the Nd_{ij} documents for the j^{th} topic of type i with a fuzziness degree comprised between f_{min} and f_{max} can be done in a natural way by (a) recursively considering all theme sub-categories of the given topic in the constructed tree $\mathcal{T}$ (see previous section), (b) randomly selecting all documents in the subcategory with at least f_{min} and at most f_{max} different topics till Nd_{ij} documents are collected, and (c) adding to the collected corpus the documents which belong to topics different from topics of types different from i. A question however arises when doing so when $f_{max} > 1$: for a given document, should one keep all the topics it belongs to or should one just disregard the ones different from the original j^{th} topic of type i? Disregarding such topics can be detrimental to our purpose of constructing gold standard clustered bilingual comparable corpora as one may lose valuable information relating documents (through disregarded topics) which are finally placed in different topics while being strongly related. We thus propose here to keep them, referring to them as *secondary* topics, and consider them as new clusters of type i. The final set of clusters thus comprises both primary and secondary topics, the latter being added to the former when collecting documents. Because of this addition, Nt_1, Nt_2 and $Nt_{1\&2}$ correspond to lower bounds of the actual number of clusters obtained, as illustrated in Table 2. However, as one can note, in most cases, as the number of topics per documents is limited in Wikipedia, one ends

ID	Language pair	Doc.	# of clusters	# of primary topics	T/D	D/T	f_{min}	f_{max}	Order	# of clusters per type	# of documents per cluster
№01	De-Fr	6982	49	43	1.19	170.12	1	5	$(2, 1\&2, 1)$ †	{10, 15, 20}	{200, 500, 750}
№02	Fr-Sw	6811	63	49	1.52	164.51	1	5	$(2, 1\&2, 1)$ †	{10, 15, 20}	{200, 500, 750}
№03	De-En	4415	33	31	1.20	160.48	1	10	$(2, 1\&2, 1)$ †	{10, 15, 20}	{100, 150, 250, 500}
№04	Fr-Ru	3386	35	32	1.16	111.80	1	10	$(2, 1\&2, 1)$ †	{10, 15, 20}	{100, 150, 250, 500}
№05	Fr-Ru	5636	20	20	1.00	281.80	1	1	$(2, 1\&2, 1)$ †	{5, 10}	{100, 150, 250, 500}
№06	En-De	5139	19	19	1.00	270.47	1	1	$(2, 1\&2, 1)$ †	{5, 10}	{100, 150, 250, 500}
№07	En-Fr	2726	18	17	1.06	160.06	1	10	$(1, 2, 1\&2)$ †	{10, 15, 20}	{100, 150, 200, 250}
№08	En-Fr	3255	21	19	1.14	176.10	1	10	$(1\&2, 2, 1)$ †	{10, 15, 20}	{100, 150, 200, 250}
№09	En-Fr	2578	17	15	1.26	190.35	1	10	$(2, 1\&2, 1)$ †	{10, 15, 20}	{100, 150, 200, 250}
№10	En-Fr	2677	122	34	2.15	47.25	2	10	$(2, 1\&2, 1)$ †	{10, 15, 20}	{100, 150, 200, 250}
№11	En-Fr	3466	24	22	1.14	164.04	1	100	$(2, 1\&2, 1)$ †	{10, 15, 20}	{100, 150, 200, 250}
№12	En-Fr	3411	17	17	1.00	200.65	1	1	$(2, 1\&2, 1)$ †	{10, 15, 20}	{100, 150, 200, 250}
№13	En-Fr	14617	31	31	1.00	471.52	1	1	$(1, 1\&2, 2)$	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
№14	En-Fr	25813	119	71	1.73	374.82	1	10	$(1, 2, 1\&2)$	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
№15	En-Fr	6460	212	21	2.98	90.92	2	10	$(1, 2, 1\&2)$	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
№16	En-Fr	13544	70	63	1.13	218.74	1	10	$(2, 1\&2, 1)$	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
№17	En-Fr	20106	94	60	1.48	315.89	1	10	$(1, 2, 1\&2)$ †	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
№18	En-Fr	30932	113	57	2.00	547.71	1	10	$(2, 1\&2, 1)$ †	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}

Table 2: This table provides details for comparable corpus, including the language pair, number of documents, number of primary topics, overall number of topics, average number of topics per document (T/D), and average number of documents per topic (D/T). The section on the right-hand side provides information on creating a corpus. This includes the minimum and maximum number of topics in documents, the order in which topic types are collected, and the range for randomly selecting the number of topics of each type and the number of documents in a topic. A special sign (†) means alternating order, while no sign means consideration by type. The level of reference l_r is 2 for all corpora.

up with a number of clusters relatively close to the original number set by the user.

In the process described above, in accordance with our will to construct clustered corpora in which documents only belong to clusters of the same type, one has to check, for every selected document, whether it belongs to clusters of different types or not. If this verification is simple, it raises the question of the ordering in which the different types of clusters are considered. Indeed, the more versatile topics, *i.e.*, the topics being commonly assigned with other topics, are more likely to be encountered at the beginning of the above process than at its end. Such topics also impact the fuzziness degree as they are likely to be present in the documents selected. We thus allow the user to play with possible orderings, firstly by deciding in which order the different types should be considered[1], and secondly by deciding to either process all topics in a given type before moving to the other types, or alternate between types after each topic. These different configurations are also illustrated in Table 2.

4. Results & Discussion

This section presents the results of our study, which consists of two main components. Firstly, technical details about the tool used and its application are provided. Secondly, the bilingual comparable corpora created with the tool are analysed to identify whether control over the number of clusters represented in only one language or both languages, the fuzziness, and the ability to adapt the corpora to a particular domain were achieved.

4.1. Tool

The code was implemented in Python (v.3.8.10), using requests (v.2.27.1), beautifulsoup4 (v.4.10.0), numpy (v.1.21.6) libraries and is freely available [2]. Information from the Wikipedia pages was obtained through MediaWiki API [3]. The tool has three functions: creating a category tree, creating a clustered bilingual comparable corpus, and visualising an obtained corpus. During the creation of the category tree, two adjustable parameters are available: a root category and the level of reference l_r used for selecting topics and thus clusters. When initiating corpus creation, several parameters can be considered, including the language pair for the collection, the range of topics present in the documents through the parameters f_{min} and f_{max}, the order in which topics are considered, the number of topics of each type, and the number of documents in each cluster. The last two parameters may either be a specific number or a set of values from which one value will be randomly selected. Additional details can be found on the code repository.

4.2. Collected Corpora

Creating a comparable corpus using Wikipedia as a source enables the development of topic-specific

[1] There are six possible choices for that: $(1\&2, 1, 2)$, $(1\&2, 2, 1)$, $(1, 1\&2, 2)$, $(1, 2, 1\&2)$, $(2, 1\&2, 1)$, and $(2, 1, 1\&2)$.

[2] https://github.com/anna-laskina/comparable_corpora_generator

[3] https://www.mediawiki.org/wiki/API:Main_page

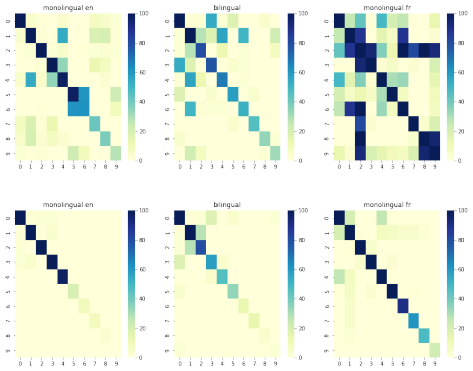

Figure 2: Heat map of the number of documents shared by the 10 largest topics of type monolingual En (left), bilingual (middle), and monolingual Fr (right) for corpus №10. Top: primary and secondary topics; bottom: primary topics only.

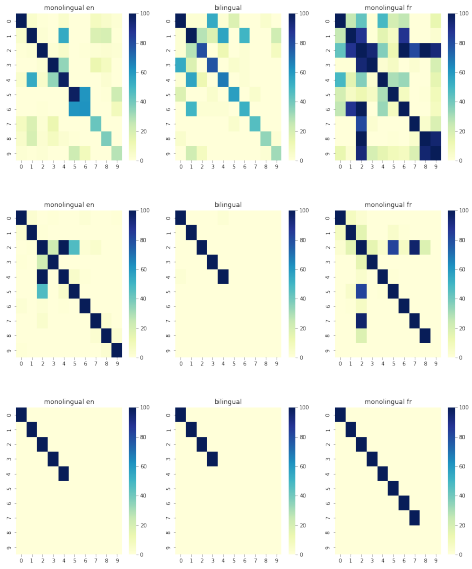

Figure 3: Heat map of the number of documents shared by the 10 largest topics of type monolingual En, bilingual, and monolingual Fr (from right to left respectively) across corpora №10, №11, №12 (from top to bottom), run with the same parameters, except f_{max}, which is equal to 10, 100, 1 for these corpora respectively, and f_{min}, which is equal to 1 for corpus №11 and №12, and equal to 2 for corpus №10.

corpora, particularly in languages with a substantial representation in Wikipedia, such as English, German, French, and Swedish. This study focuses on the English-French language pair, but also includes corpora for other language pairs such as English-German, French-Russian, French-German, and French-Swedish. A series of corpora were generated to analyse the effectiveness of the corpus generation algorithm. In this paper, we provide detailed descriptions of 18 corpora, with pertinent information delineated in Table 2.

There are two types of obtained topics: primary and secondary. Primary topics are those initially selected when the collection began, while secondary topics are those that appeared during the collection process when an article with unreferenced topics was added to the collection; these topics were added as new clusters and acquired the type as the requested topic. Considering the top 10 topics of each type reveals that secondary topics exhibit a greater dispersion of documents beyond the main diagonal (Fig. 2). This observation suggests that primary topics tend to be more coherent, with fewer documents containing multiple primary topics. In contrast, secondary topics introduce fuzziness into documents, facilitating a higher incidence of multiple topics within a single document.

Subsequently, we examined three datasets initialized with different values of f_{min} and f_{max} alongside consistent remaining parameters to discern the varying degrees of fuzziness attainable. Fixing f_{max} at 1 yields completely non-fuzzy (hard) clustering, depicted in Figure 3 (bottom). Conversely, selecting f_{min} at 2 and f_{max} at 10 facilitates achieving fuzzy clustering, as illustrated in Figure 3 (top). The mean number of topics per document across corpora ranges from 1.00 to 2.98 (Table 2), whereas within a single corpus, the number of topics per document can reach the maximum value defined by f_{max} (see Figure 4).

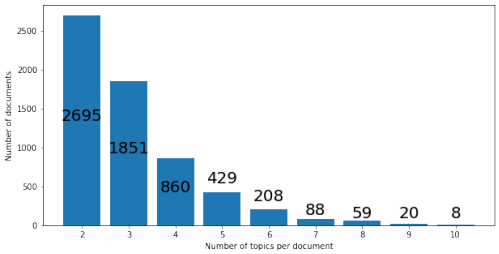

Figure 4: The distribution of documents in corpus №15 by the number of topics present within it.

When executing corpus collection with parameters varying solely in the order of topic consideration, it becomes evident that when topics are arranged by type style (as depicted by $(1, 2, 1\&2)$ on the top and $(2, 1\&2, 1)$ in the middle of Figure 5), fuzziness becomes concentrated in the monolingual ℓ_1 and monolingual ℓ_2 categories, respectively, as they were the first types considered. Conversely, when topics are arranged by alternating style (bottom of Figure 5), fuzziness is more evenly distributed across different types. However, achieving precise control over the localization of the fuzzier

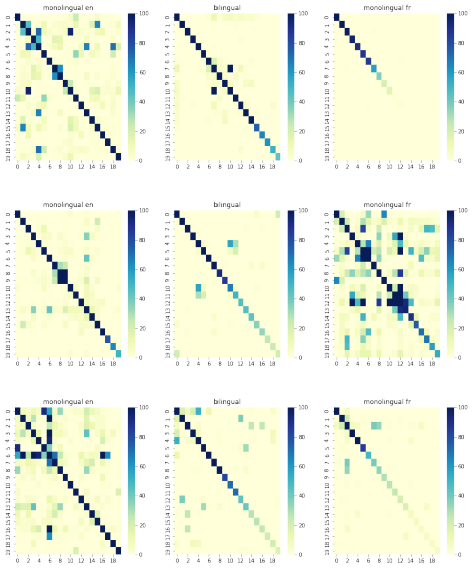

Figure 5: Heat map of the number of documents shared by the 20 largest topics of type monolingual En, bilingual, and monolingual Fr (from right to left respectively) across corpora №14, №16, №17 (from top to bottom), run with the same parameters, except the order in which topic types are considered: $(1, 2, 1\&2)$ *by type* for corpus №14, $(1, 2, 1\&2)$ *alternating* for corpus №17 and $(2, 1\&2, 1)$ *by type* for corpus №16.

segment becomes more challenging, as the second type of topics gains little advantage from being considered earlier than the final third type. Additionally, although the style of order in consideration influences the distribution of topics by types, a more significant correlation is observed initially from the selection of specific topics for each type.

Finally, customization of the category tree creation according to preferences is feasible. Ones have the option to select the root category and a set of topics for corpora. In our context, the category *Main topic classifications* was selected as the root category, as we did not have specific topic preferences. However, one can narrow the selected cluster to a particular area and choose, for example, the *Health* category as the root category (Fig. 6). The selection of the level of reference l_r in the obtained tree allows one to further focus on specific subcategories of, *e.g.*, the Health domain.

5. Conclusion

We have presented in this paper a method to extract clustered bilingual comparable corpora from Wikipedia with different fuzziness levels and language representativity. Wikipedia is an excellent source for constructing such corpora because of its categorised articles and interlingual links, which

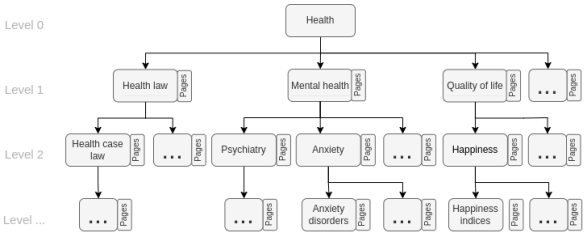

Figure 6: The category tree with the category *Health* as root.

facilitate the creation of bilingual links between articles. After extracting the topical tree backbone of the Wikipedia category system, we have proposed a construction process which allows one to somehow regulate the fuzziness level (*i.e.*, the fact that a document can be associated with more than one cluster) of the obtained corpus, as well the representativity of each language. Indeed, clustered bilingual comparable corpora are characterized by the fact that they contain three types of clusters: those consisting of documents in either language only, and those comprising documents from the two languages.

Our analysis has shown that it is possible to exert considerable influence over the above corpus characteristics, achieving significant control over fuzziness levels and language representativity, as well as determining the subject domain of the corpus. Future enhancements of the proposed methodology could include the method of collecting Wikipedia corpora on a particular topic proposed by España-Bonet et al. (2023). We also plan to extend our tool to directly construct clustered comparable corpora in more than two languages.

6. Acknowledgements

This work has been funded by the French projects ANR-20-IDES-0005 IDéES@UGA and ANR-19-P3IA-0003 MIAI@Grenoble Alpes.

7. Bibliographical References

Ahmad Aghaebrahimian, Andy Stauder, and Michael Ustaszewski. 2022. Testing the validity of wikipedia categories for subject matter labelling of open-domain corpus data. *Journal of Information Science*, 48(5):686–700.

Iñaki Alegria, Unai Cabezon, Unai Fernandez de Betono, Gorka Labaka, Aingeru Mayor, Kepa Sarasola, and Arkaitz Zubiaga. 2013. Reciprocal enrichment between basque wikipedia and machine translation. *The People's Web Meets*

NLP: Collaboratively Constructed Language Resources, pages 101–118.

Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, and Malek Ezzeddine. 2016. Derivation of "is a" taxonomy from wikipedia category graph. *Engineering Applications of Artificial Intelligence*, 50:265–286.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Nicolas Aspert, Volodymyr Miz, Benjamin Ricaud, and Pierre Vandergheynst. 2019. A graph-structured dataset for wikipedia research. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1188–1193.

Alberto Barrón-Cedeno, Cristina España Bonet, Josu Boldoba Trapote, and Luís Márquez Villodre. 2015. A factory of comparable corpora from wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. 2021. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In *Computational Linguistics and Intelligent Text Processing*, pages 296–309, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, page 380–392, Berlin, Heidelberg. Springer-Verlag.

Cristina España-Bonet, Alberto Barrón-Cedeño, and Lluís Màrquez. 2023. Tailoring and evaluating the wikipedia for in-domain comparable corpora extraction. *Knowledge and Information Systems*, 65(3):1365–1397.

Vishal Goyal, Ajit Kumar, and Manpreet Singh Lehal. 2020. Document alignment for generation of english-punjabi comparable corpora from wikipedia. *International Journal of E-Adoption (IJEA)*, 12(1):42–51.

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*, pages 291–300.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

P Otero, I López, S Cilenis, and Santiago de Compostela. 2011. Measuring comparability of multilingual corpora extracted from wikipedia. *Iberian Cross-Language Natural Language Processings Tasks (ICL)*, page 8.

Pablo Gamallo Otero and Isaac González 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.

Magdalena Plamada and Martin Volk. 2012. Towards a wikipedia-extracted alpine corpus.

Magdalena Plamadă and Martin Volk. 2013. Mining for domain-specific parallel text from Wikipedia.

In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 112–120, Sofia, Bulgaria. Association for Computational Linguistics.

Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. 2018. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Motaz Saad, David Langlois, and Kamel Smaïli. 2013. Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia-Social and Behavioral Sciences*, 95:40–47.

Krzysztof Suchecki, Alkim Almila Akdag Salah, Cheng Gao, and Andrea Scharnhorst. 2012. Evolution of wikipedia's category structure. *Advances in complex systems*, 15(supp01):1250068.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Kun Yu and Jun'ichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. In *Proceedings of Machine Translation Summit XII: Posters*.

Torsten Zesch and Iryna Gurevych. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 1–8, Rochester, NY, USA. Association for Computational Linguistics.