BACK TO THE ROOTS: TRACING SOURCE LANGUAGES IN WIKIPEDIA WITH LABSE

Eleonora Cupin, Lucia Galiero, Debora Ciminari
Departament of Translation and Interpretation
University of Bologna, Forlì

Abstract: Our objective is to try and ascertain the plausible source language of Wikipedia articles on two domains, gastronomy and science, across five languages: Italian, English, Spanish, German, and French. Drawing upon the concept of text similarity, we explore its application in multilingual Wikipedia pages, which can serve as a potential source for the creation of comparable and parallel corpora, though they do not always represent direct translation of one another. The choice fell on the abovementioned domains so that we can examine the feasibility of discerning the source language in a cultural and non-cultural field. For this purpose, this study attempts to estimate text similarity by computing the cosine similarity between sentence embeddings generated by the LaBSE model.

Key Words: *Text similarity, LaBSE, Wikipedia, Cosine Similarity, Parallel sentence mining, NLP*

1 Introduction

This study aims to explore the notion of text similarity, which can be defined as the commonality between two text snippets, with similarity being greater as the commonality increases [12]. In particular, we consider text similarity on a semantic level rather than a morphological and syntactic one. In general, this notion represents a major focus within a wide range of Natural Language Processing (NLP) applications, such as information retrieval, machine translation, and automatic question answering [20]. By drawing on such a concept, this study aims to infer the plausible source language of Wikipedia articles written in five languages (Italian, English, Spanish, German, and French) and discussing the same topics from two domains (gastronomy and science). By addressing both a cultural field and a scientific field, this study tries to examine what the plausible source language is. Wikipedia is the object of this study since it provides articles on the same topic in different languages (i.e., interlanguage-linked articles), thus constituting a source of extensive multilingual data and one of the largest and most popular sources of comparable data for training machine translation systems [2]. Nevertheless, Barrón Cedeño et al. (2014) point out that, while some articles are mutual translations, others are independent of one another[2]. The possibility for users to edit Wikipedia pages entails a multi-authored content production, which brings about articles' inherent instability [14].

Our work's main contribution is thus applying a multilingual approach to the task of identifying the plausible source language for a set of sentences sampled from a comparable corpus, extracted from Wikipedia. For this purpose, we explore sentence-level text similarity by attempting to identify plausible parallel sentences across five languages. To this end, we use sentence embeddings, generated through the Language-agnostic BERT Sentence Embedding (LaBSE) model [5]. In order to assess the text similarity of each pair of sentences, we compute the cosine similarity between their representations.

Additionally, we carry out a manual evaluation on the same 200 sentences (100 for each domain) in three out of the five languages (Italian, English, Spanish) to determine how many pairs of sentences correspond to parallel sentences. Our evaluation is then analyzed by assessing the inter-

annotator agreement through Krippendorff's alpha to estimate the level of agreement among the annotators.

2 Related work

Among the studies addressing text similarity, Fujishiro et al. (2023) develops a semantic search system for a Japanese database collecting medical malpractice claims [6]. They employ Sentence-BERT (SBERT) to create the embeddings and use the Euclidean distance to assess the similarity between pairs of embeddings. Yet Fujishiro et al. (2023)'s study only approaches the problem in a monolingual setting, focusing on Japanese. Other studies address more than one language and attempt to select parts in comparable data. [1][16] These studies, however, are developed in the context of domain adaptation, which does not ensure that the selected parts are parallel, i.e., source and target sentences are in-domain, but they are not necessarily translations [19]. Moreover, corpora created in such a way are not suitable to train Neural Machine Translation (NMT) systems because they are sensitive to noise [19].

3 Data[1]

As described in Figure 1, the data used in this study is divided into two categories. The first dataset comes from the GeNTE corpus, compiled by Fondazione Bruno Kessler [17]. It consists of 1,500 parallel sentences extracted from the English-Italian and English-Spanish segments of the Europarl corpus, which is based on the proceedings of the European Parliament in the official languages of the EU [10]. This was chosen because of its high parallel quality, which ruled out potential grammatical or translation inconsistencies, and the higher probability that the content was written by a human. Also, this ensures that our methodology is tested in a relatively predictable environment, providing a foundation for the identification of a cosine similarity threshold as accurate as possible. The second dataset is further divided into two domains, i.e., gastronomy and science. Overall, it comprises a total of 105 selected Wikipedia entries for each of the two domains (210 main entries in total), ensuring that each was available in all the 5 languages selected for the present study, for a total of 1,050 texts. The aim of this division is to test which observations regarding the possible source language can be made across both domains, i.e., whether results might be generalized in a cross-domain perspective or whether some discerning might be required when dealing with culture-specific topics. As regards the gastronomic domain, each of the five languages had to be represented with an equal number of entries on cultural-specific dishes. This means that, out of 105 dedicated main entries, 21 had to belong to Italian cuisine, 21 to the French, and so on. With relation to the scientific domain, such a clear-cut division into equally represented subdomains was not made, assuming that the scientific domain and its related sub-branches are not culturally specific nor culture-dependent. Generally speaking, the sample used for scientific text includes texts on human anatomy, geology, astronomy and nuclear physics.

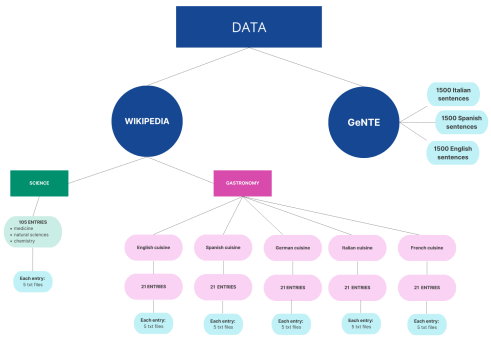

Figure 1: Data - Wikipedia and GeNTE

4 Model

Following the work of [3], we use the LaBSE (Language-agnostic BERT Sentence Embedding) model [5] to calculate sentence-level representations. LaBSE is a BERT-based model [4], trained on a corpus encompassing 109 languages which employs a dual-encoder architecture, with one encoder processing the source sentence and the other processing the target sentence. Initialized with pre-trained BERT weights, the model's encoders are trained using a translation ranking loss with an additive margin, prioritizing correct translations over incorrect ones [5], as shown in Figure

[1] https://drive.google.com/drive/folders/1KJwvjil4gszxaGmq0kso54LsvobntNxG?usp=sharing

2. In this setup, LaBSE embeddings are pivotal as they serve as the primary input for subsequent similarity assessments. Leveraging a multilingual vector space, the system generates vector representations of sentences sourced from a multilingual dataset. Each output embedding consists of a 768-dimension vector. These vectors are then compared using the cosine similarity metric to gauge the similarity between sentences.
Cosine similarity is key for assessing vector similarity, particularly in the context of comparing sentence embeddings. It has a convenient range for most machine learning problems, [-1, 1]. Vectors with a cosine similarity close to 1 leads to the conclusion that the two sentences are likely talking about the same thing [9].
Therefore, by dividing texts into sentences, and converting them into vectors, LaBSE captures their semantic essence, enabling a more nuanced and precise comparison across different languages. This method stands on the premise that true semantic similarity transcends mere textual or syntactic parallelism, reaching into the realm of conceptual equivalence [15].

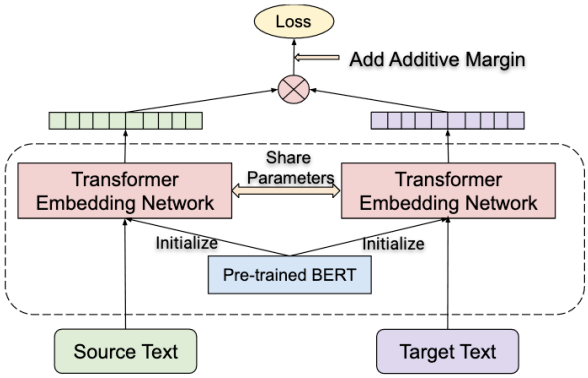

Figure 2: LaBSE model graph.

5 Methodology[2]

Our methodology for detecting the source language in the collected data is founded upon a sequential use of Python scripts, each tailored to a specific objective. Execution of these scripts occurs primarily on Google Colab and Visual Studio Code platforms. Due to the substantial volume of data under analysis, reliance solely on open-source CPU resources proved inadequate. In Colab, PyTorch was configured to use the cuda hardware architecture, enabling GPU utilization and leveraging its complimentary availability until such resources were no longer accessible. Subsequently, a transition to the computing infrastructure provided by the Department of Translation and Interpretation (Forlì) was done.

5.1 Threshold

As introduced in Section 3, the GeNTE corpus [17] was used to determine a threshold value for cosine similarity aimed at classifying sentences as either parallel or non-parallel. A similar methodology was applied by Feng et al. (2022), who investigated parallel text mining from the CommonCrawl dataset [5]. They employed a binary classification approach based on an arbitrary cosine similarity threshold of 0.6. In their methodology, sentence pairs scoring above this threshold were categorized as parallel, while those below it were deemed non-parallel. However, given that the CommonCrawl dataset collects raw webpage data, it was decided to assess LaBSE's performance first on highly parallel texts, specifically reliable translations, before making any decisions. The model yielded the following averages, calculated on sentence similarities of all fragments contained in the corpus:

5.2 Getting data from Wikipedia

After obtaining the threshold value, the next step involved automating the retrieval and archiving of Wikipedia pages. To achieve this, the 'wikipedia' library was used. The script is based primarily on a function, i.e. 'download and save page', designed to interact efficiently with Wikipedia's API. This function simplifies the process of fetching Wikipedia page content in a specified language while addressing potential issues such as page unavailability or disambiguation.
In order to correctly use this function, the Wikipedia page titles are written within a dictionary with their respective language. The key of this dictionary, instead, represents the name of the subfolder that will be created and saved within the main folders that represent our domains either "Gastronomy" or "Science". As result each subfolder contains 5 txt files, with the language

[2]https://drive.google.com/drive/folders/1$_{CojNHQJobWA1TfX2GXAfkleNwdwV9qK?usp=sharing}$

appended to their filenames.

5.3 Document cleaning

Since Wikipedia texts frequently feature numerous sections, such as notes, bibliographic references, redirections to other entries and links to external websites, the following step focused on removing such information, which could potentially add noise across the dataset. After identifying the specific patterns to be removed, the cleaning stage was carried out by making use of the regular expressions in Python to remove all titles in general, avoiding that similarity between couples of noun phrases would compromise results by adding sets of phrases with similarity scores too close or equal to 1.

5.4 Embedding computation and similarity calculation

Each text was split into sentences with NLTK's sentencizer. Sentences were processed with a batch size of 8 due to VRAM constraints, and tokenized using the BERT tokenizer. All the tokenized data was then passed to LaBSE to compute embeddings for each sentence in each language for subsequent similarity comparison.
We performed padding to ensure that all sentences in each batch had the same number of tokens. Each sentence with less than the maximum number of tokens for the given batch will be padded until the sentence length reaches that said value.
Moreover, we had to ensure that the number of sentences in each set of documents were the same as well. For this reason, the pipeline integrates the creation of zero tensors that were concatenated to the sentence embeddings, uniforming the number of sentences for comparison of articles in different languages. This was designed to facilitate similarity calculations and ensure that sentences without a candidate equivalent could still be aligned without affecting the final results.
Results were compiled into frames, 210 in total, with each being dedicated to one single entry (e.g. The document "Anemia.csv" contains all cosine similarities for all 5 Wikipedia articles on anaemia in different languages). Every single frame displays sentence pairs in different language combinations, with the respective language abbreviations and the calculated cosine similarity for the couple of sentences at stake. Once we obtained all similarity scores, we proceeded to filter sentences at or above our established threshold (0.75 - see Section 5.1.).
Finally, the last portion of the pipeline is dedicated to infer the potential source language (being defined by us as the language with the highest number of sentences in combination with the others) of articles in our given (sub)domains. This was done by summing up the occurrences of each individual language within all the calculated cosine similarities. The results were summarized in a dataframe for each "cuisine" in the case of the gastronomy domain, whereas for the scientific domain a single data frame was used.

6 Manual evaluation[3]

Having obtained the results from the LaBSE model, we aimed to observe and identify potential translation pairs. To conduct this evaluation, we randomly selected a sample of 200 sentences, 100 from the science domain and 100 from the gastronomy domain, and performed a manual assessment of sentences with a similarity score equal to or above the chosen threshold of 0.75. Excel was used as the platform for this manual evaluation, and the data were assessed by the three authors of this study. The languages included in the sample were English, Italian, and Spanish, as these were the common languages among the three annotators.
The chosen evaluation scale ranged from 1 to 4: 1) completely different, 2) almost completely different, 3) highly similar, 4) parallel translation. This scale was selected to avoid intermediate values and encourage the annotators to make more definitive judgments. To prevent bias, the annotators were asked to hide both the column with the cosine similarities and the columns with the scores given by the other annotators. After having manually evaluated the sentences, the level of agreement between the annotators was measured using Krippendorff's alpha, a statistical measure of inter-rater reliability that assesses the consistency of ratings given by multiple observers [11].

[3] https://drive.google.com/drive/folders/14oQM0u55PUsBx0HUJZ2MhVp-mXIWbPRN?usp=sharing

7 Results and discussions

7.1 LaBSE cosine similarity

As shown in Figure 3, we can more confidently assume that entries collected from Wikipedia in the domain of gastronomy, which are culture-specific, tend to originate in the language of the culture they pertain to. However, the final calculation, which considered the number of cosine similarities where each individual language was present, revealed some additional insights.
For instance, within German gastronomy, French (497) and German (556) show only slight differences. A similar observation can be made between Italian (825) and English (924) in the context of English gastronomy, and between Italian (601) and French (588) in the context of Italian gastronomy. A distinct case is Spanish gastronomy, which significantly stands out from the other languages with a result of 839.
Regarding texts in the scientific domain, the thesis supported by Scarpa et al. (2015) is confirmed, as English emerges as the predominant and therefore pivot language.

Gastronomy — GERMAN CUISINE

	EN	IT	FR	DE	ES
EN	0	65	115	104	78
IT	65	0	73	141	76
FR	115	73	0	217	92
DE	104	141	217	0	94
ES	78	76	92	94	0

The prevailing language is 'DE' with a total count of 556

Gastronomy — ENGLISH CUISINE

	EN	IT	FR	DE	ES
EN	0	411	213	78	222
IT	411	0	145	68	201
FR	213	145	0	58	145
DE	78	68	58	0	43
ES	222	201	145	43	0

The prevailing language is 'EN' with a total count of 924

Gastronomy — SPANISH CUISINE

	EN	IT	FR	DE	ES
EN	0	74	118	114	259
IT	74	0	71	61	203
FR	118	71	0	75	238
DE	114	61	75	0	139
ES	259	203	238	139	0

The prevailing language is 'ES' with a total count of 839

Gastronomy — FRENCH CUISINE

	EN	IT	FR	DE	ES
EN	0	66	45	53	75
IT	66	0	107	47	61
FR	45	107	0	40	109
DE	53	47	40	0	47
ES	75	61	109	47	0

The prevailing language is 'FR' with a total count of 301

Gastronomy — ITALIAN CUISINE

	EN	IT	FR	DE	ES
EN	0	179	150	51	106
IT	179	0	264	37	121
FR	150	264	0	49	125
DE	51	37	49	0	36
ES	106	121	125	36	0

The prevailing language is 'IT' with a total count of 601

Science — MEDICINE, CHEMISTRY, NATURAL SCIENCES

	EN	IT	FR	DE	ES
EN	0	4164	3018	1285	3425
IT	4164	0	2468	1567	2998
FR	3018	2468	0	1450	2152
DE	1285	1567	1450	0	1125
ES	3425	2998	2152	1125	0

The prevailing language is 'EN' with a total count of 11892

Figure 3: Dataframes with the similarities.

Wikipedia, being a tool for dissemination, naturally includes notes and bibliographical citations. In academic and scientific fields, researchers often prefer to write articles in English for publication in specialized journals. Additionally, many university-level textbooks are typically translations of texts originally written in English [18]. These same sources are some of those cited or rewritten by the authors of individual Wikipedia entries, making it unsurprising that the information reported likely had an English source.

7.2 Manual Evaluation

The obtained Krippendorff's alpha coefficient among the three annotators is 0.52, indicating moderate agreement among the annotations (Figure 4). This suggests that there is some consistency but there may also be discrepancies due to various factors, such as improving the guidelines provided or the presence of noise in the data.
It was observed that there were few sentences not correctly segmented. Our segmentation process relied on the NLTK's sentencizer and there could be the possibility that whether the space after the period was missing, the split did not occur.
Another aspect observed is that a sentence in one language appeared very similar with two sentences in another language. Considering this was an experimental evaluation, a coefficient of 0.52 is deemed quite satisfactory, especially given that the noise accounts for only 10 percent of all sentences.
Regarding the number of sentences identified as parallel by the annotators, approximately 60 sentences out of 200 were flagged with a score of 4 (i.e, parallel translations). The average cosine similarity of these sentences was 0.94, confirming the estimates made by the LaBSE model at the beginning of the present study.

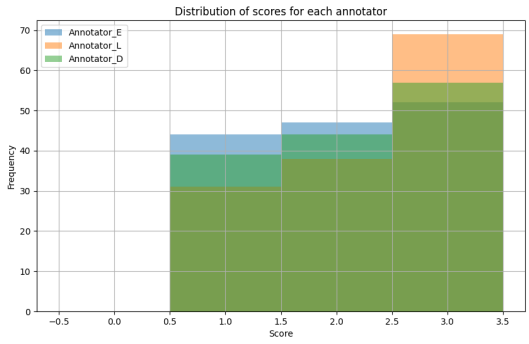

Figure 4: Bar chart of annotators' agreement.

8 Limitations and future research

Given the results previously illustrated, it could be argued that it is not always possible to clearly discern the source language since the nature of Wikipedia articles defies the traditional directionality of translation. Specifically, within the domain of science, English emerges as the source language, while in the more cultural sub-domains of gastronomy, a source language does not distinctly prevail. As a result, it was only feasible to state that the filtered sentences were comparable or parallel, whereas distinguishing between source and target was not always attainable.
Future work might test our methodology on more than five languages as well as under-resourced languages, which may be limited by a smaller number of entries. Moreover, possible additions could adopt other models, such as LASER (Language-Agnostic Sentence Embeddings Representations [3]). Finally, this study can contribute to exploring how semantic-based text matching works when using sentence embeddings in a multilingual setting and can provide insights on semantic search, i.e., an information retrieval process which is based on the meaning behind user queries rather than keyword matching [13]. Modern search engines are shifting away from a traditional keyword-based approach to develop more semantically informed engines with an enhanced understanding of online content [7]. Among its advantages, semantic search is crucial because users do not always use the same language as the desired result [13].

Bibliography

[1] Axelrod A., He X. an Gao J.,(2011), Domain adaptation via pseudo in-domain data selection. In EMNLP, pp. 355–362.

[2] Barrón C. A., Paramita ML.; Clough P.; Rosso P. (2014). A Comparison of Approaches for Measuring Cross-Lingual Similarity of Wikipedia Articles. En Advances in Information Retrieval. Springer Verlag (Germany). pp. 424-429. doi:10.1007/978-3-319-06028-636.

[3] Chimoto E., Bassett B., (2022). Very Low Resource Sentence Alignment: Luhya and Swahili. In Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022), Gyeongju, Republic of Korea. Association for Computational Linguistics, pp. 1-8.

[4] Devlin J.,(2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in J. Burstein, C. Doran, and T. Solorio (eds) 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). NAACLHLT 2019, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. Available at: https://doi.org/10.18653/v1/N19-1423.

[5] Feng F.,(2022), Language-agnostic BERT Sentence Embedding, arXiv. Available at: https://doi.org/10.48550/arXiv.2007.01852.

[6] Fujishiro N., Otaki Y., Kawachi S., (2013), Accuracy of the Sentence-BERT Semantic Search System for a Japanese Database of Closed Medical Malpractice Claims. https://doi.org/10.3390/app13064051

[7] Giomelakis D., (2023), Semantic Search Engine Optimization in the News Media Industry: Challenges and Impact on Media Outlets and Journalism Practice in Greece. Social Media + Society. 9.

[8] Gurevych and Y. Miyao, (2018), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia: Association for Computational Linguistics, pp. 228–234. Available at: https://doi.org/10.18653/v1/P18-2037.

[9] Hobson L., Cole H., Hannes H. (2019). Natural Language Processing in Action Understanding, analyzing, and generating text with Python. Manning Publications.

[10] Koehn P., (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, In Proceedings of Machine Translation Summit X: Papers, Phuket, Thailand, pp. 79–86.

[11] Krippendorff K., (2011), Computing Krippendorff's Alpha-Reliability.

[12] Lin D., (1998), An information-theoretic definition of similarity, In Proceedings of the International Conference on Machine Learning, Madison, WI, USA, pp. 296–304.

[13] Pecánek M., (2020), What is semantic search? How it impacts SEO. Ahrefs. https://ahrefs.com/blog/semantic-search/

[14] Ray A., Graeff E., (2008), Reviewing the Author-Function in the Age of Wikipedia, In C. Eisner M. Vicinus (Eds.), Originality, Imitation, and Plagiarism: Teaching Writing in the Digital Age, University of Michigan Press, pp. 39–47. https://doi.org/10.2307/j.ctv65sxk1.6

[15] Redzioch-Korkuz, A. Revisiting the concepts of translation studies equivalence in linguistic translation from the point of view of Peircean universal categories" Language and Semiotic Studies, vol. 9, no. 1, 2023, pp. 33-53.

[16] Santamaría L., Axelrod A., (2017). Data selection with cluster-based language difference models and cynical selection. In IWSLT, pp. 137–145.

[17] Savoldi B., Piergentili A., Fucci D., Negri M., Bentivogli L., (2024), "A Prompt Response to the Demand for Automatic Gender-Neutral Translation". In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, Malta.

[18] Scarpa F., (2015), L'influsso dell'inglese sulle lingue speciali dell'italiano, International Journal of Translation n.16, EUT Edizioni, Università di Trieste, Trieste, pp. 225-243.

[19] Schwenk H., (2018), Filtering and Mining Parallel Data in a Joint Multilingual Space, In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia. Association for Computational Linguistics, pp. 228–234.

[20] Wang J, Dong Y., (2020), Measurement of Text Similarity: A Survey. Information., https://doi.org/10.3390/info11090421.