
Proceedings Paper:
Paramita, M.L. orcid.org/0000-0002-9414-1853, Clough, P. and Gaizauskas, R. (2017) Using Section Headings to Compute Cross-Lingual Similarity of Wikipedia Articles. In: Jose, J.M., Hauff, C., Altıngovde, I.S., Song, D., Albakour, D., Watt, S. and Tait, J., (eds.) ECIR 2017: Advances in Information Retrieval. 39th European Conference on Information Retrieval (ECIR 2017), 08/04/2017 - 13/04/2017, Aberdeen, UK. Lecture Notes in Computer Science (10193). Springer, Cham , pp. 663-669. ISBN 978-3-319-56608-5

https://doi.org/10.1007/978-3-319-56608-5_59

Using Section Headings to Compute Cross-Lingual Similarity of Wikipedia Articles

Monica Lestari Paramita[1], Paul Clough[1] and Robert Gaizauskas[2]

[1] Information School, University of Sheffield, UK
[2] Computer Science Department, University of Sheffield, UK
{m.paramita,p.d.clough,r.gaizauskas}@sheffield.ac.uk

Abstract. Measuring the similarity of interlanguage-linked Wikipedia articles often requires the use of suitable language resources (e.g., dictionaries and MT systems) which can be problematic for languages with limited or poor translation resources. The size of Wikipedia can also present computational demands when computing similarity. This paper presents a 'lightweight' approach to measure cross-lingual similarity in Wikipedia using section headings rather than the entire Wikipedia article, and language resources derived from Wikipedia and Wiktionary to perform translation. Using an existing dataset we evaluate the approach for 7 language pairs. Results show that the performance using section headings is comparable to using all article content, dictionaries derived from Wikipedia and Wiktionary are sufficient to compute cross-lingual similarity and combinations of features can further improve results.

Keywords: Wikipedia similarity, cross-language similarity

1 Introduction

As the largest Web-based encyclopedia, Wikipedia contains millions of articles written in 295 languages and covering a large number of domains[3]. Many articles describe the same topic in different languages, connected via *interlanguage-links*. Measuring cross-lingual similarity within these articles is required for tasks, such as building comparable corpora [4]. However, this can be challenging due to the large number of Wikipedia language pairs and the limited availability of suitable language resources for some languages [7]. Language-independent methods for computing cross-lingual similarity have been proposed, for example based on character n-gram overlap, but the accuracy of such methods decreases significantly for dissimilar language pairs [2].

Based on previous work [6], we propose a method for computing similarity across languages using scalable, yet lightweight, approaches based on structural similarity (comparing section headings) and using translation resources built from Wikipedia and Wiktionary. This paper addresses the following research questions: *(RQ1) How effective are section headings for computing article similarity compared to using the full content?* and *(RQ2) How effective is information derived from Wikipedia and Wiktionary for translating section headings compared to using high-quality translation resources?*

[3] https://en.wikipedia.org/wiki/List_of_Wikipedias (20 Oct 2016)

2 Related Work

Since interlanguage-linked Wikipedia articles describe the same topic, they have often been assumed to contain similar content and have been utilised for various tasks, such as mining parallel sentences [1] and building bilingual dictionaries [3]. The similarity of these articles across languages, however, may vary widely and have not been thoroughly investigated in the past. One study that analysed Wikipedia similarity [6] identified characteristics contributing to cross-lingual similarity, including overlapping named entities and similar structure. Features, such as the overlap of links, character n-gram overlap and cognate overlap of the article contents have been investigated as ways to automatically identify cross-lingual similarity with promising results [2]. Previous work, however, have not explored structural similarity features to identify cross-lingual similarity of Wikipedia articles.

The approach we propose makes use of Wikipedia and Wiktionary to assist in translating section headings (previously identified as a possible indicator of article's structural similarity [6]), prior to computing similarity. Both resources have been used to compute cross-lingual similarity [1, 5] and semantic relatedness [8]. However, past work has often focused on highly-resourced language pairs. This study investigates the use of these resources for under-resourced language pairs.

3 Methodology

The content of most Wikipedia articles are structured into sections and subsections, e.g. the Wikipedia article of "United Kingdom" includes the following section headings (titles): *Etymology, History, Geography,* etc. Our method aims to measure cross-lingual similarity between a document pair D_1 and D_2 in a non-English language (L_1) and English (L_2) by measuring the similarity between their section headings, which is computationally more efficient than comparing entire contents. We refer to these section headings as H_1 and H_2, respectively. The approach is described in Section 3.1 and evaluation setup in Section 3.2.

3.1 Proposed Approach to Compute Cross-Lingual Similarity

Dictionary Creation. Firstly, two dictionaries are built using Wikipedia and Wiktionary, a multilingual dictionary available in 152 languages[4]. An existing link-based bilingual lexicon method [1] was used to extract the titles of Wikipedia interlanguage-linked articles for each language pair, using them as dictionary entries. We supplemented this lexicon with entries from Wiktionary, as this contains more lexical knowledge compared to Wikipedia [5]. This was performed by collecting English Wiktionary entries and their translations in non-English language pairs.

[4] https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries (20 Oct 2016).

Translation of Section Headings. Firstly, common headings that do not make useful contributions when computing article similarity, such as *References, External Links* and *See Also*, were filtered out. Stopwords were also removed using a list of frequent words gathered from Wikipedia (an average size of 871 words per language). Afterwards, the English section headings (H_2) are translated into L_1 (the non-English language), resulting in H_2'. For each section heading (h_1, h_2, ..., h_n) in H_2, the translation process is as follows:

1. If h_i exists in the dictionary, then extract all of its translations t_i.
2. If h_i does not exist as an entry in the dictionary:
   (a) If h_i includes > 1 word, split the heading h_i into each word (w_1, w_2, ..., w_n) and translate each word separately.
   (b) If no translation is found for a given word, trim 1 character from the end of the word and search for its translation. Perform this recursively until either a translation is found, or the original word has 4 characters left.
   (c) Perform step (a) for all words in h_i and concatenate the results.
3. Both h_i and t_i (if found) are then included in H_2'.
4. Steps 1-3 are repeated until all headings in H_2 have been translated.

Identification of Structural Similarity. In this stage, we aim to align similar section headings in both documents. Firstly, every source heading $s_i \in H_1$ is paired to every target heading $t_j \in H_2'$. For each s_i, we identify the most similar target heading t_n (allowing many-to-one alignments) using the following alignment and section similarity scoring (*secSimScore*) methods:

1. If s_i is contained in t_j, both headings are aligned; $secSimScore(s_i, t_j) = 1$.
2. If not, split heading s_i into each word $(w_1, w_2, ..., w_p)$:
   (a) Find if w_m is included in t_j. If not, recursively trim w_m by 1 character until either it is included in t_j, or w_m has 4 characters left.
   (b) Perform step (a) for all words in s_i; $secSimScore(s_i, t_j)$ is calculated by measuring the proportion of words in s_i that are found in t_j.
3. Step 1-2 are performed between s_i and the remaining sections in H_2'. After which, the highest scoring pair is selected as the alignment for s_i.

After all the aligned sections in H_1 and H_2' are identified, referred to as A_1 and A_2', respectively ($A_1 \in H_1$ and $A_2' \in H_2'$), the scores are aggregated to derive a structure similarity score for the document pair (*docSimScore*). Three different methods to measure the *docSimScore* are investigated:

1. **align1**: This method does not take the *secSimScore* of the aligned sections into account, but instead relies on the number of aligned sections in both documents only:

$$docSimScore = \frac{(|A_1| + |A_2'|)}{(|H_1| + |H_2'|)} \tag{1}$$

   where $|A_1|$ and $|A_2'|$ represent the number of aligned sections in H_1 and H_2', respectively, and $|H_1|$ and $|H_2'|$ are the number of sections in H_1 and H_2'.

2. **align2**: This method takes the $secSimScore$ into account. In Equation 1, $|A_1|$ is replaced with the sum of $secSimScore$ for each aligned section in A_1.
3. **align3**: In this method, aligned sections with $secSimScore < 1$ are filtered out, prior to calculating $align3$ using Equation 1.

An additional feature, *the ratio of section length* (**sl**), is also extracted by measuring the ratio of number of section headings in both articles.

3.2 Evaluation Setup

To evaluate the approach we used an existing Wikipedia similarity corpus [6] containing 800 document pairs from 8 language pairs. Two annotators assessed the similarity of each document pair using a 5-point Likert Scale. Due to the unavailability of Wiktionary translation resource in Croatian-English, only 7 language pairs are used in this study: German (a *highly-resourced* language), and 6 *under-resourced* languages: Greek (EL), Estonian (ET), Lithuanian (LT), Latvian (LV), Romanian (RO) and Slovenian (SL); all paired to English (EN). Documents without section headings were removed for these experiments, resulting in 600 document pairs across the 7 language pairs. We compare the proposed methods to **c3g**, the tf-idf cosine similarity of the *char-3-gram overlap* between the article contents[5]. To investigate the effectiveness of Wikipedia-Wiktionary as translation resources, we use Google Translate as a state-of-the-art comparison.

4 Results and Discussion

(RQ1) How effective are section headings for computing article similarity compared to using the full content? We report the Spearman-rank correlations between similarity scores computed using methods from Section 3.1 and the average human-annotated similarity scores from the evaluation corpus in Table 1 ("Individual Features"). Results show that features based on section headings (ρ=0.36 for $align1$) were able to achieve comparable overall correlations compared to using char-3-gram overlap ($c3g$) on the entire article contents (ρ=0.34). Results using $align2$ was similar (ρ=0.35). The $align3$ method, however, achieved significantly lower score (ρ=0.23), suggesting that the strict alignment process may have lost valuable cross-lingual information. Section length (sl) was shown to perform consistently across most language pairs (ρ=0.35). The $c3g$ method, however, performed poorly for RO-EN and SL-EN (ρ=0.20 and ρ=0.03, not statistically significant), possibly due to dissimilar surface forms between languages. Section heading features were shown to achieve either the same or better correlation scores than $c3g$ in 5 of the 7 language pairs.

Our findings also suggest that a combination of features produces a more robust similarity measure. Table 1 ("Combined Features") reports the three best feature combinations. Firstly, a combination of only Section Headings (SH)

[5] This feature was previously identified as the best language-independent feature to identify cross-lingual similarity in Wikipedia [2].

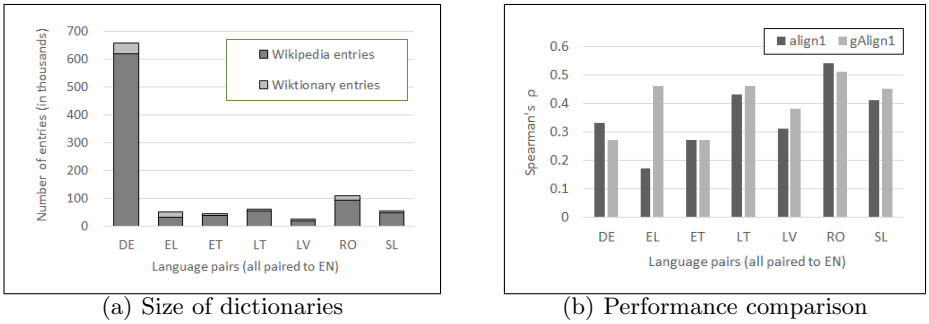
Table 1: Correlation scores (Spearman's ρ) of individual and combined features

Lang	Individual Features					Combined Features		
	Section Headings (SH)				Article	SH	SH + Article	
	align1	**align2**	**align3**	**sl**	**c3g**	**align1_sl**	**sl_c3g**	**align1_sl_c3g**
DE	0.33*	0.28	-0.01	0.45*	**0.46***	0.42*	**0.67***	0.59*
EL	0.17	0.19	0.19	**0.42***	0.38*	0.36*	**0.56***	0.47*
ET	0.27*	0.29*	0.29*	0.37*	**0.57***	0.37*	**0.58***	0.54*
LT	0.43*	**0.44***	0.39*	0.40*	0.34*	0.54*	0.51*	**0.58***
LV	0.31*	0.33*	0.18	**0.34***	**0.34***	0.40*	0.46*	**0.49***
RO	**0.54***	**0.54***	0.51*	0.14	0.20	**0.40***	0.20	0.39*
SL	**0.41***	0.32*	0.00	0.33*	0.03	**0.44***	0.33*	0.42*
Avg	**0.36**	0.35	0.23	0.35	0.34	0.42	0.49	**0.50**

Note: $*p < 0.01$; the best results for the "Individual Features" and "Combined Features" are shown in bold; "Avg" score is calculated using Fisher transformation.

features, *align*1_*sl*, increases the correlation score to 0.42 (↑16.67% compared to *align*1, the best individual feature). Correlation can further be increased by combining both SH and article features. We show that *sl_c3g* achieves ρ=0.49 (↑36.11%); considering that this feature can be computed without the need of a dictionary, this result is very promising. Lastly, the combination of three features, *align*1_*sl_c3g*, achieves the highest correlation score (ρ=0.50; ↑38.89%).

(RQ2) How effective is information derived from Wikipedia and Wiktionary for translating section headings compared to using high-quality translation resources? Figure 1(a) shows the dictionary size derived from Wikipedia and Wiktionary used in this study, highlighting low numbers of entries for all under-resourced languages. To investigate the effect of different translation resources, we computed the *align*1 method using a high-quality translation resource: in this case Google Translate (*gAlign*1). The correlation scores of the original *align*1 method (using the Wiki resources) and *gAlign*1 are shown in Figure 1(b)). Although a much higher *gAlign*1 correlation was achieved in EL-EN (ρ=0.46, compared to ρ=0.17 for *align*1), the correlation scores for the remain-



(a) Size of dictionaries (b) Performance comparison

Fig. 1: Translation Resources

ing language pairs are very similar. In some language pairs (DE-EN, ET-EN, and RO-EN), the use of Wikipedia-Wiktionary resources achieved either the same or better correlation scores compared to using Google Translate. Our findings also show that the dictionary size does not significantly affect the performance of the section heading alignment methods. For example, LV-EN, which has the smallest dictionary (24.4K entries) achieves similar $align1$ correlation to DE-EN (the largest dictionary with 641K entries). We also found that, although much smaller in size, an average of 66% of Wiktionary entries are not available in the Wikipedia lexicon; this shows the importance of Wiktionary in complementing the Wikipedia lexicon.

5 Conclusions and Future Work

This paper describes a 'lightweight' approach for identifying cross-lingual similarity of Wikipedia articles by measuring the structural similarity (i.e. similarity of section headings) of the articles. Results show that the section heading similarity feature ($align1$) and ratio of section length (sl) can be used to identify cross-lingual similarity with comparable performance to using the overlap of char-3-grams ($c3g$) on content from the entire article (ρ=0.36, ρ=0.35, and ρ=0.34, respectively). A combination of these three features also further improves the results (ρ=0.50). The use of Wikipedia-Wiktionary resource in this approach was shown to be as efficient to utilising Google Translate for many language pairs. These results are promising as these resources are freely available for a large number of languages. Future work will investigate more feature combinations and to measure similarity in Wikipedia in more language pairs.

References

1. Adafre, S.F., de Rijke, M.: Finding similar sentences across multiple languages in Wikipedia. In: Proceedings of EACL '06. pp. 62–69 (4 Apr 2006)
2. Barrón-Cedeño, A., Paramita, M.L., Clough, P., Rosso, P.: A comparison of approaches for measuring cross-lingual similarity of Wikipedia articles. In: Proceedings of ECIR '14. pp. 424–429. ECIR '14, Springer (2014)
3. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: An approach for extracting bilingual terminology from Wikipedia. In: Database Systems for Advanced Applications, LNCS, vol. 4947, pp. 380–392. Springer Berlin Heidelberg (2008)
4. Mohammadi, M., GhasemAghaee, N.: Building bilingual parallel corpora based on Wikipedia. In: ICCEA 2010. vol. 2, pp. 264–268. IEEE (2010)
5. Müller, C., Gurevych, I.: Using Wikipedia and Wiktionary in domain-specific information retrieval. In: Proceedings of CLEF 2008. pp. 219–226 (2009)
6. Paramita, M.L., Clough, P., Aker, A., Gaizauskas, R.J.: Correlation between similarity measures for inter-language linked Wikipedia articles. In: LREC '12. pp. 790–797 (2012)
7. Yasuda, K., Sumita, E.: Method for building sentence-aligned corpus from Wikipedia. In: Proceedings of WikiAI '08. pp. 64–66 (13–14 Jul 2008)
8. Zesch, T., Müller, C., Gurevych, I.: Using wiktionary for computing semantic relatedness. In: AAAI Conference on Artificial Intelligence. pp. 861–866 (2008)