A Comparison of Approaches for Measuring Cross-Lingual Similarity of Wikipedia Articles

Alberto Barrón-Cedeño[1], Monica Lestari Paramita[2],
Paul Clough[2], and Paolo Rosso[3]

[1] Talp Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
albarron@lsi.upc.edu http://www.lsi.upc.edu/∼albarron
[2] Information School, University of Sheffield, Sheffield, UK
[m.paramita,p.d.clough]@sheffield.ac.uk http://ir.shef.ac.uk/
[3] NLE Lab, PRHLT, Universitat Politècnica de València, Valencia, Spain
prosso@dsic.upv.es http://users.dsic.upv.es/∼prosso

Abstract. Wikipedia has been used as a source of comparable texts for a range of tasks, such as Statistical Machine Translation and Cross-Language Information Retrieval. Articles written in different languages on the same topic are often connected through inter-language-links. However, the extent to which these articles are similar is highly variable and this may impact on the use of Wikipedia as a comparable resource. In this paper we compare various language-independent methods for measuring cross-lingual similarity: character n-grams, cognateness, word count ratio, and an approach based on outlinks. These approaches are compared against a baseline utilising MT resources. Measures are also compared to human judgements of similarity using a manually created resource containing 700 pairs of Wikipedia articles (in 7 language pairs). Results indicate that a combination of language-independent models (char-n-grams, outlinks and word-count ratio) is highly effective for identifying cross-lingual similarity and performs comparably to language-dependent models (translation and monolingual analysis).

Key words: Wikipedia · Cross-Lingual Similarity

1 Introduction

Wikipedia, the free online encyclopedia, contains articles on a diverse range of topics written by multiple authors worldwide. It has been used as a source of multilingual data for a range of monolingual and cross-language NLP and IR tasks, such as named entity recognition [14], query translation for CLIR [9], word-sense disambiguation [6] and statistical machine translation [7]. Wikipedia editions have been created in more than 287 languages, with some languages more evolved than others. A proportion of topics appear in multiple Wikipedias (i.e. in multiple languages), resembling a comparable corpus. However, the similarity and textual relationship between articles written in multiple languages on the same topic (further referred to as *interlanguage-linked articles*) can vary widely. Some articles may be translations of each other; others may have been written independently and cover different aspects of the same topic. In some cases, the articles may even contain contradictory information [3].

Measuring similarity is core to many tasks in IR and NLP. However, few studies have focused on computing Cross-Lingual (CL) similarity in Wikipedia; particularly for under-resourced languages. In this work, we are interested in methods that require few, if any, language resources ("language-independent" approaches). This is important in cases where languages being studied are "under-resourced". We identify various models to compute cross-lingual similarity in Wikipedia (Section 2) and use them to calculate cross-language similarity between 700 Wikipedia article pairs (Section 3). We measure performances of these models using an existing Wikipedia evaluation benchmark [10]. We conclude the paper in Section 4 and provide directions for future research.

2 Cross-Lingual Similarity Models

Multiple models to assess the degree of similarity across languages have been proposed for different tasks, such as the extraction of parallel text fragments for MT [8] and CLIR [2, 4]. The purpose is to estimate a similarity value $sim(d, d')$, where $d \in L$ $d' \in L'$ are words, text fragments, or documents written in languages L, L' ($L \neq L'$). The process for CL similarity estimation involves three steps. (i) *Pre-processing*: d and d' are passed by a standard normalisation process (e.g. tokenisation, case folding, etc.). (ii) *Characterisation*: d and d' are mapped into a common space —some strategies are based on MT techniques, translating d into L', allowing for further monolingual analyses; others break the texts into small chunks (e.g. character n-grams) and exploit syntactic similarities between languages [5, 12]; yet other techniques attempt to map concepts in d and d' into a common semantic space on the basis of multilingual thesauri [13] or comparable corpora [11]. (iii) *Comparison*: The mappings of d and d' are compared on the basis of a similarity measure (e.g. cosine). In this paper, we focus on a range of methods requiring few, if any, specific language resources. Other models, such as CL-ESA [11], exhibit state-of-the-art performance; however, because we are also interested in methods that will operate efficiently and for under-resourced languages, we do not consider such resource-demanding approaches. To evaluate performance we compare methods against a language-dependent method which utilises MT followed by a monolingual analysis.

Character *n*-grams (c*n*g). To calculate char-n-grams, a simplified alphabet $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$ is considered; i.e. any other symbol, space, and diacritic is discarded and case-folding applied. The text is then codified into a vector of character n-grams ($n = [3, 5]$). This model is an adaptation of McNamee & Mayfield's [5].

Cognateness (cog). This concept was proposed in [12] to identify parallel sentences. A token t forms a *cognateness* candidate if: (a) t contains at least one digit; (b) t contains only letters and $|t| \geq 4$; or (c) t is a punctuation mark. t and t' are pseudo-cognates if both belong to (a) or (b) and are identical, or belong to (b) and share the same four leading characters. Hence, we characterise d and d' as follows: if t accomplishes (a) , it is maintained verbatim, if it accomplishes (b) it is cut down to its first four characters. We neglect (c) because we are comparing entire articles. Again, case-folding and removal of diacritics is applied.

Word Count Ratio (wc). This simple measure is computed as the length ratio between the shorter and the longer document (in number of tokens).

Common Outlinks (lnk). This is a model appropriate for analysing Wikipedia articles that has been used on the extraction of similar sentences across languages with encouraging results [1]. It exploits the Wikipedia's internal *outlinks*: if an article in language L (L') links to article a_L ($b_{L'}$) and a and b are about the same topic, the corresponding texts are considered similar. We compute a simplified version where a vector is created representing outlinks in the documents that are mapped to another language using the structure of Wikipedia.

Translation + Monolingual Analysis (trans$_n$). Our baseline is a language-dependent model: we use Google's MT system to translate d into L', generating d_t, which are then compared against d' using a standard monolingual process.

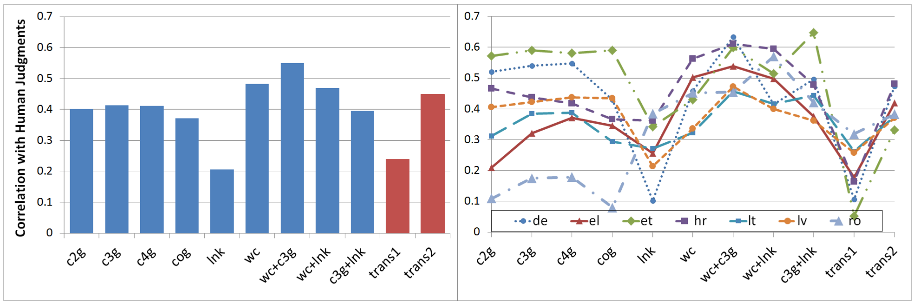
3 Experiments

We selected 7 language pairs: German (de), Greek (el), Estonian (et), Croatian (hr), Latvian (lv), Lithuanian (lt), and Romanian (ro), which are all paired to English (en). These languages were chosen as they exhibit different characteristics, such as writing systems (Greek texts were transliterated using ICU4J: http://site.icu-project.org) and availability of resources: German is highly resourced; the remaining languages are under-resourced.

First, we calculate similarity scores for each document pair using the models described in Section 2. We use an existing Wikipedia evaluation benchmark [10] containing 100 document pairs for each language pair, ranging between 107-1,546 words.[4] The similarities were manually scored between 1–5 by two assessors with appropriate language skills. Overall, 85% of the pairs were given the same score or differ by one, which shows a good agreement between the assessors. For each document pair, we compute the manually-assessed similarity score by averaging assessors' scores, resulting in 9 similarity values (ranging from 1-5). Lastly, we compare the models by calculating Spearman-rank correlation between (combined) automatic to human-assessed similarity scores.

Spearman-rank correlation coefficient scores are calculated between the automatically- and manually-generated similarity scores for all 700 document pairs

[4] We discard Slovenian due to the high number of pairs judged as similar in the evaluation set (>95%) as it would affect the accurate calculation of correlation scores.



(a) All languages

(b) Per language

Fig. 1. Correlation of different models and human judgements. Note: cng=character n-grams (n=[2-4]); cog=cognateness; lnk=outlinks; wc=word counts; trans$_n$ =translation plus word n-grams comparison.

and each similarity method (cf. Figure 1(a)). Language-independent models such as char-n-grams ('c2g', 'c3g' and 'c4g') identify cross-lingual similarity with performance comparable to the baseline translation models using bi-gram overlap ('trans$_2$'). The results show that a simplistic language-independent model based on the word count ratio ('wc') correlates higher with human judgements compared to models using MT, which suggests that interlanguage-linked Wikipedia articles with similar lengths are very likely to contain similar content. This correlation, however, can still be improved by using a combination of syntax-based (specifically 'c3g') and structure-based models ('wc' or 'outlinks'). These findings are promising, considering that these models are purely language-independent and can easily be calculated for many language pairs.

Whilst language-independent models perform well overall, their performance may differ for each language pair. Therefore, we also computed correlations in each language. Figure 1(b) shows that whilst char-n-grams perform well on average, their correlations vary significantly across different languages. The simplified outlinks ('lnk') model was less reliable in identifying similarity. However, combining 'lnk' and 'wc' results in a more stable model, performing well for all languages, although slightly lower than the combination of 'wc' and 'c3g'. This combination obtains ρ=0.55; just slightly lower than the combination of 'wc' and 'trans$_2$': ρ=0.57 (not shown in the plot).

We perform a similar analysis for language-dependent models and identify a drastic increase in correlation of using word bi-grams ('trans$_2$') compared to unigrams ('trans$_1$') overlap. The poor performance for the latter may also be caused by the weighting strategy used: simple tf. A straightforward enhancement of this model would be to remove stopwords and apply tf-idf weighting.

Table 1 shows the various models and their correlation scores (top-performing models highlighted). For all language pairs, the highest correlation to human judgements are achieved using combinations of language-independent methods. For five language pairs, 'wc+c3g' is proven to be superior, while 'wc+lnk' and

Table 1. Average correlation scores across document pairs for each language pair.

	Lang-independent									Lang-dependent	
	Syntax-based				Struct-based		Combinations			MT-based	
lan	**c2g**	**c3g**	**c4g**	**cog**	**lnk**	**wc**	**wc+c3g**	**wc+lnk**	**c3g+lnk**	**trans$_1$**	**trans$_2$**
de	0.52	0.54	0.55	0.43	0.10	0.46	**0.63**	0.41	0.50	0.11	0.47
el	0.21	0.32	0.37	0.35	0.26	0.50	**0.54**	0.50	0.38	0.18	0.42
et	0.57	0.59	0.58	0.59	0.34	0.43	0.60	0.52	**0.65**	0.05	0.33
hr	0.47	0.44	0.42	0.37	0.36	0.56	**0.61**	0.59	0.48	0.16	0.48
lt	0.31	0.38	0.39	0.29	0.27	0.32	**0.46**	0.42	0.44	0.26	0.38
lv	0.41	0.42	0.44	0.43	0.21	0.34	**0.47**	0.40	0.36	0.26	0.37
ro	0.11	0.17	0.18	0.08	0.38	0.45	0.45	**0.57**	0.42	0.32	0.38

'c3g+lnk' perform better in Romanian–English and Estonian–English, respectively. These results are promising: by combining language-independent models, it is possible to reliably identify cross-lingual similarity in Wikipedia with better performance (i.e. correlation to human judgement) than using MT systems.

4 Conclusions

This paper compares different methods for computing cross-lingual similarity in Wikipedia. Methods vary from language-independent (syntax-based, structure-based, and a combination of the two), to language-dependent requiring language-specific resources, e.g. an MT system. In contrast to previous work, we investigated the performance of each method for a wide range of under-resourced languages. We analysed correlations between these models and human judgements by making use of an existing evaluation benchmark for Wikipedia.

We conclude that a combination of language-independent models perform better than language-dependent models (i.e. involving translation) of bi-gram word overlap. Word count ratio and char-3-grams perform best in most languages ($\rho = 0.55$), followed by a combination of word count and outlinks ($\rho = 0.47$). A simple translation model using word bi-gram correlates with manual judgements (ρ=0.45), followed by the last method combination: char-n-grams and outlinks (ρ=0.40). This result is very promising given that these models can be calculated without the need of any translation resources, which will enable these models to be applied to measure cross-lingual similarity to any Wikipedia language pair, possibly after applying transliteration.

As future work we plan to investigate other combinations of language-independent models in order to create a language-independent approach to reliably measure cross-lingual similarity in Wikipedia. We also plan to use these models to further investigate similarity between articles in Wikipedia.

Acknowledgements

The work of the first author was in the framework of the Tacardi research project (TIN2012-38523-C02-00). The work of the fourth author was in the framework of the

DIANA-Applications (TIN2012-38603-C02-01) and WIQ-EI IRSES (FP7 Marie Curie No. 269180) research projects.

References

1. Adafre, S., de Rijke, M.: Finding Similar Sentences across Multiple Languages in Wikipedia. In: Proc. of the 11th Conf. of the European Chapter of the Association for Computational Linguistics. pp. 62–69 (2006)
2. Dumais, S., Letsche, T., Littman, M., Landauer, T.: Automatic Cross-Language Retrieval Using Latent Semantic Indexing. In: AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval. pp. 24—26. Stanford University (1997)
3. Elena, F.: Directions for exploiting asymmetries in multilingual Wikipedia. In: Proc. of the Third Intl. Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies. Boulder, CO (2009)
4. Levow, G.A., Oard, D., Resnik, P.: Dictionary-Based Techniques for Cross-Language Information Retrieval. Information Processing and Management: Special Issue on Cross-Language Information Retrieval 41(3), 523–547 (2005)
5. Mcnamee, P., Mayfield, J.: Character N-Gram Tokenization for European Language Text Retrieval. Information Retrieval 7(1-2), 73–97 (2004)
6. Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation. In: Proc. of NAACL 2007. ACL, Rochester (2007)
7. Mohammadi, M., GhasemAghaee, N.: Building Bilingual Parallel Corpora based on Wikipedia. In: Second Intl. Conf. on Computer Engineering and Applications. vol. 2, pp. 264–268 (2010)
8. Munteanu, D., Fraser, A., Marcu, D.: Improved Machine Translation Performace via Parallel Sentence Extraction from Comparable Corpora. In: Proc. of the Human Language Technology and North American Association for Computational Linguistics Conf. (HLT/NAACL 2004). Boston, MA (2004)
9. Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D., Hiemstra, D., de Jong, F.: WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia. Proc. of the Cross-Language Evaluation Forum LNCS (5706) (2009), springer-Verlag
10. Paramita, M.L., Clough, P.D., Aker, A., Gaizauskas, R.: Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In: Calzolari, e.a. (ed.) Proc. of the 8th Intl. Language Resources and Evaluation (LREC 2012). pp. 790—797. ELRA, Istanbul, Turkey (2012)
11. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-Based Multilingual Retrieval Model. Advances in Information Retrieval, 30th European Conf. on IR Research LNCS (4956), 522–530 (2008), springer-Verlag
12. Simard, M., Foster, G.F., Isabelle, P.: Using Cognates to Align Sentences in Bilingual Corpora. In: Proc. of the Fourth Intl. Conf. on Theoretical and Methodological Issues in Machine Translation (1992)
13. Steinberger, R., Pouliquen, B., Hagman, J.: Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. Computational Linguistics and Intelligent Text Processing. Proc. of the CICLing 2002 LNCS (2276), 415—424 (2002), springer-Verlag
14. Toral, A., Muñoz, R.: A proposal to automatically build and maintain gazetteers for Named Entity Recognition using Wikipedia. In: Proc. of the EACL Workshop on New Text 2006. Association for Computational Linguistics, Trento, Italy (2006)