Pairing Wikipedia Articles Across Languages

Marcus Klang
Lund University
Department of Computer Science
Lund, Sweden
Marcus.Klang@cs.lth.se

Pierre Nugues
Lund University
Department of Computer Science
Lund, Sweden
Pierre.Nugues@cs.lth.se

Abstract

Wikipedia has become a reference knowledge source for scores of NLP applications. One of its invaluable features lies in its multilingual nature, where articles on a same entity or concept can have from one to more than 200 different versions. The interlinking of language versions in Wikipedia has undergone a major renewal with the advent of Wikidata, a unified scheme to identify entities and their properties using unique numbers. However, as the interlinking is still manually carried out by thousands of editors across the globe, errors may creep in the assignment of entities. In this paper, we describe an optimization technique to match automatically language versions of articles, and hence entities, that is only based on bags of words and anchors. We created a dataset of all the articles on persons we extracted from Wikipedia in six languages: English, French, German, Russian, Spanish, and Swedish. We report a correct match of at least 94.3% on each pair.

1 Introduction

Wikipedia has become a major reference knowledge source for applications such as IBM Watson (Ferrucci, 2012) or Google's knowledge graph (Singhal, 2012). Wikipedia is available in more than 280 languages such as Indonesian, Amharic, Nahuatl, or Hindi and its content and coverage are continuously growing thanks to millions of contributors.

One of the major steps to organize the linguistic diversity of Wikipedia has been the creation of Wikidata: A centralized repository of entities and concepts identified by unique numbers. Before Wikidata, an editor creating an article, say on Madagascar in Spanish, had to link it manually to already existing versions of the same entity. Now, Wikidata stores links to all the versions in a centralized repository (Fig. 1) and adding a new language is carried out through this repository and a unique number: Q1019 in the case of Madagascar. In addition to associating unique identifiers to entities, Wikidata uses a set of about 2,500 properties, as of June 1, 2016, to describe them. One of these properties is *instance of*, P31, that enables the editors to define an ontology. Madagascar, for example, is an instance of a sovereign state, an island, an island nation, a country, and a member state of the United Nations (Fig. 1).

Although Wikidata has simplified the linking process, it is still a manual operation that is not error-free and articles may be incorrectly linked across the languages. In this paper, we report and evaluate a technique that, given a human entity, automatically identifies the set of articles describing this entity across six Wikipedia languages: English, French, German, Russian, Spanish, and Swedish. We report a correct match of at least 94.7% for each pair selected among the six versions.

2 Previous Work

Comparable corpora, like the language versions of Wikipedia, have been used extensively as resources to extract word translations or parallel sentences. Rapp et al. (2012) for instance, used Wikipedia articles in nine languages to identify word translations through keywords and a word alignment algorithm. Schamoni et al. (2014) proposed to use links to retrieve Wikipedia articles in English similar to an article

Proceedings of the Open Knowledge Base and Question Answering (OKBQA) Workshop,
pages 72–76, Osaka, Japan, December 11 2016.

Figure 1: **Left part**: The first language versions of Madagascar in Wikidata. The languages appear in alphabetic order out of 195. **Right part**: Membership of Madagascar to ontology classes using the *instance of* property; three classes are listed out of five

in German. Domínguez García et al. (2012) extracted hyponymy relations from the Wikipedia category system across languages. Chiao and Zweigenbaum (2002) used a set of documents in French and English collected with medical taxonomy terms to produce translational equivalents. Sproat et al. (2006) used English and Chinese stories from the Xinhua News agency to identify named entity transliterations. Finally, Smith et al. (2010) improved translation performance using sets of parallel sentences that they extracted from Wikipedia.

Although these works carry out some kind of matching across languages, we could not find references on a systematic attempt to pair descriptions of an entity in multiple language versions. To the best of our knowledge, we are the first to propose and evaluate a method in this field.

3 Collecting the Corpus

We used six dumps of Wikipedia in English, French, German, Russian, Spanish, and Swedish[1] from which we extracted all the persons. We carried out this extraction using the *instance of* property, where we collected all the articles that had a direct link to the node denoting a human entity (Q5 in Wikidata). Table 1, left part, shows the counts of articles on persons broken down per language.

From this person data set, we extracted all the articles having the six language versions. Again, we used the Wikidata identifier to determine the available language versions of an entity. We obtained a total of 1,938,861 unique persons having at least one version in one of the six languages and where 39,636 had versions in the six languages (Table 1, right part).

Language	Count	Language	Count	Versions	Count	Versions	Count
en	1,257,604	ru	297,202	6	39,636	3	115,692
de	579,656	es	262,538	5	42,986	2	284,658
fr	446,308	sv	178,894	4	65,725	1	1,390,164

Table 1: **Left part**: Counts of articles on persons per language version. **Right part**: Number of versions for the articles on persons

4 Method

To implement the comparison method, we restricted the articles to their first paragraphs that we represented as bags of words and entity identifiers (Wikidata Q-number). For a given pair of languages, we then determined the best pair of paragraphs using the cosine similarity.

The articles on persons in Wikipedia show a similar structure, where the first paragraph starts with the name of the person and reports a few basic facts such as the dates and places of birth and death using

[1] Retrieved in May and September 2015.

numbers and proper nouns. In languages using the Gregorian calendar, the numbers are often the same across the versions and many proper nouns also have identical forms and spellings.

For instance, the first paragraph on Shakespeare in the French Wikipedia,

> **William Shakespeare**, né probablement le **26** avril **1564** à Stratford-upon-**Avon** et mort le 3 mai (**23** avril) **1616** dans la même ville, [...] à représenter les aspects de la nature humaine.

shares seven words with the corresponding paragraph in English:

> **William Shakespeare** (/ˈʃeɪkspɪər/; **26** April **1564** (baptised) – **23** April **1616**) was an English poet, playwright, and actor, [...] and the "Bard of **Avon**" [...] than those of any other playwright.

while it does not share a single word with Napoléon's one:

> Napoléon Bonaparte (/nəˈpoʊliən, -ˈpoʊljən/; French: [napɔleɔ̃ bɔnapaʁt], born Napoleone di Buonaparte; 15 August 1769 – 5 May 1821) was a French military and political leader [...] Napoleon implemented foundational liberal reforms in France and throughout Europe. [...]

We represented each article as a bag of words of its first paragraph with the vector space model (Salton et al., 1974). In addition to the words, we used resolved anchors from the first paragraph with their Q-numbers as terms. We extracted all the unique tokens and Q-numbers from all the documents from which we excluded 250 stop words that we defined as the words that occur in the highest number of documents across the four versions. We used this variant of $TF \cdot IDF$: $weight(\text{term}) = tf(\text{term}, d) \cdot \log \frac{N}{df(\text{term}, D)}$, where tf is the term frequency in the current document d (a paragraph); df is the document frequency, that is the number of paragraphs D that contain this term; N is the total number of documents; in our case, the total number of paragraphs.

Given a pair of languages, the two sets of articles in their respective languages and their association form a weighted bipartite graph, where the comparison (matching) step can be formulated as a linear assignment problem (Jacobi, 1865; Kuhn, 1955; Jonker and Volgenant, 1987). The worst case of computing the weight matrix involves $O(N^2)$ operations, which for our dataset corresponds to 1.57 billion operations, while the assignment problem has a $O(N^3)$ worst case complexity. This figure is still in the realm of feasibility, but could quickly get worse with more categories. Fortunately, our comparisons involve pairs that typically share few terms, and most often none at all. In this case, their cosine similarity is 0. In our data set, 95% of the matrix elements are zero, which makes the computation of an optimal solution tractable using a sparse linear assignment algorithm.

5 Exploratory Analysis

Taking advantage of the sparsity, we conducted an exploratory analysis and a preliminary evaluation. We compared paragraphs that shared at least one term and we implemented a simplified assignment algorithm that reduced drastically the number of operations. Given a pair of languages, the source and the target, we compared each document from the source with all the target documents that shared at least one term with it and we assigned the target document that had the maximal similarity. This simplified assignment corresponds to the initial cover in Kuhn (1955).

We applied the comparison algorithm to the set of articles. For all the language pairs, the number of misclassified articles is less than 17.0%, a surprisingly low figure for such a simple method. Table 2 shows the results, where the most confused language pair is French–English and the less one, Swedish–German.

We also examined the influence of the cosine similarity on the method using the precision and recall scores. We applied a cutoff to this similarity to validate a pair that we varied between 0 and 1. Figure 2 shows the recall and precision on the Swedish–French pair with respect to this cutoff. All the other pairs show a similar pattern. A cutoff of 0 always selects the highest cosine similarity whatever its value, while a 1 will request the paragraphs to have exactly the same words. We can see that a very high recall is reached without cutoff, while the precision is moderately improved by it. A perfect precision is reached when the cosine similarity is greater than 0.76, while a high cutoff discards all the pairs.

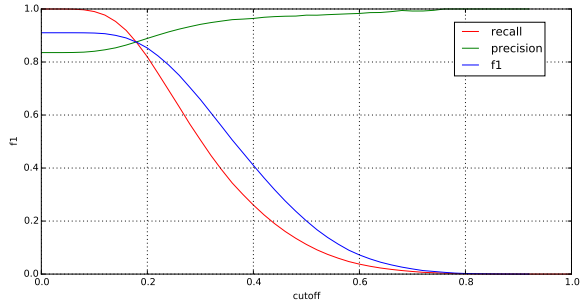

Figure 2: Recall and precision with respect to the cutoff for the Swedish–English pair

	de	en	es	fr	ru	sv
de	–	15.5	8.6	6.6	9.7	6.3
en	11.8	–	13.6	14.8	14.9	13.9
es	6.0	14.5	–	8.4	12.8	8.4
fr	5.3	17.0	9.9	–	12.7	8.7
ru	7.1	14.8	11.4	10.3	–	10.7
sv	5.0	16.4	10.0	8.7	13.4	–

Table 2: Percentages of misclassified pairs using the naïve method. First column: source language; first row: target language

	de	en	es	fr	ru	sv	avg
de	–	2.0	1.9	1.7	3.1	1.0	1.9
en	2.0	–	3.2	3.2	5.0	2.6	3.2
es	2.0	3.1	–	2.6	4.8	2.4	3.0
fr	1.7	3.2	2.6	–	4.7	2.4	2.9
ru	3.1	5.0	4.8	4.6	–	5.7	4.7
sv	0.9	2.6	2.4	2.4	4.9	–	2.7

Table 3: Percentages of misclassified pairs using the linear assignment algorithm.

6 Linear Assignment

Finally, we applied the sparse linear assignment algorithm by Jonker and Volgenant (1987) to all the pairs in our data set using a modified version of the Fiji library from Schindelin et al. (2012). For each pair, we excluded the documents that shared no word and we applied the algorithm to the resulting matrix. In spite of the matrix sizes (nearly $40,000 \times 40,000$) and the algorithm complexity, the run time to compute an assignment matrix takes less than two hours on Intel Xeon desktop computer. Table 3 shows the results, where the most confused language pair is Russian–Swedish and the less one, Swedish–German.

7 Discussion and Conclusion

The matching method we proposed shows it could reach a high accuracy with a standard bag-of-word technique, including words and entity identifiers, even with the simplified assignment method. This surprisingly high accuracy is due to the similar structure adopted by most articles on persons in their first paragraph. The proper nouns and the dates this paragraph contains proved to be sufficiently discriminative to have error rates less than 5.7% across the languages.

We applied this method to languages having the highest number of views per hour on Wikipedia[2], English and Spanish, as well as French and Swedish, that show no or little proper noun inflection, and hence where the proper nouns are identical across the versions. Nonetheless, the errors we obtained with Russian, if larger, are still comparable to those we got with the other languages: The pair (en, fr) shows an error of 3.2%, while (en, ru) is of 5.0%, for instance. Such results can be explained by the Q-numbers appearing in the bag-of-word vectors that are shared by the languages of a pair, whatever the script or morphology. They suggest this method is applicable to nonLatin scripts or to languages with a richer inflection. We believe this technique paves the way for an automatic matching of comparable textual resources across languages as well as interactive tools to support the creation and linking of new articles.

Acknowledgments

This research was supported by Vetenskapsrådet, the Swedish research council, under the *Det digitaliserade samhället* program.

[2] `http://stats.wikimedia.org/`

References

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 2*, COLING '02, pages 1–5.

Renato Domínguez García, Sebastian Schmidt, Christoph Rensing, and Ralf Steinmetz, 2012. *Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I*, chapter Automatic Taxonomy Extraction in Different Languages Using Wikipedia and Minimal Language-Specific Information, pages 42–53. Springer Berlin Heidelberg, Berlin, Heidelberg.

David Angelo Ferrucci. 2012. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4):1:1 –1:15, May-June.

Carl Jacobi. 1865. De investigando ordine systematis aequationum differentialum vulgarium cujuscunque. *Journal für die reine und angewandte Mathematik*, 64:297–320.

Roy Jonker and Anton Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Reinhard Rapp, Serge Sharoff, and Bogdan Babych. 2012. Identifying word translations from comparable documents without a seed lexicon. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 460–466, Istanbul, Turkey, May.

Gerard Salton, A. Wong, and C. S. Yang. 1974. A vector space model for automatic indexing. Technical Report TR74-218, Department of Computer Science, Cornell University, Ithaca, New York.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–494, Baltimore, Maryland, June. Association for Computational Linguistics.

Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 07.

Amit Singhal. 2012. Introducing the knowledge graph: things, not strings. Official Google Blog. `http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html`, May.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411.

Richard Sproat, Tao Tao, and ChengXiang Zhai. 2006. Named entity transliteration with comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 73–80, Sydney, Australia, July.