Identifying Multilingual Wikipedia Articles based on Cross Language Similarity and Activity

Khoi-Nguyen Tran
The Australian National University
Canberra, ACT 0200, Australia
khoi-nguyen.tran@anu.edu.au

Peter Christen
The Australian National University
Canberra, ACT 0200, Australia
peter.christen@anu.edu.au

ABSTRACT

Wikipedia is an online free and open access encyclopedia available in many languages. Wikipedia articles across over 280 languages are written by millions of editors. However, the growth of articles and their content is slowing, especially within the largest Wikipedia language: English. The stabilization of articles presents opportunities for multilingual Wikipedia editors to apply their translation skills to add articles and content to smaller Wikipedia languages. In this poster, we propose similarity and activity measures of Wikipedia articles across two languages: English and German. These measures allow us to evaluate the distribution of articles based on their knowledge coverage and their activity across languages. We show the state of Wikipedia articles as of June 2012 and discuss how these measures allow us to develop recommendation and verification models for multilingual editors to enrich articles and content in Wikipedia languages with relatively smaller knowledge coverage.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; I.5.3 [**Pattern Recognition**]: Clustering—*Similarity measures*

General Terms

Experimentation, Languages, Measurement

Keywords

Similarity, Activity, Measures, Wikipedia, Multilingual

1. INTRODUCTION

Wikipedia is the largest online free encyclopedia written by millions of volunteers and accessed by hundreds of millions of people each month. Since its introduction in 2001, its growth in article and content has been exponential. However, since 2007 growth has slowed significantly across all languages, especially English [11]. The English Wikipedia is still the largest by far in number of articles and size of articles, compared to the other 280+ languages available. While researchers determine why growth is slowing and ways to increase growth [11], we see an opportunity for Wikipedia to transform into a more complete multilingual resource.

There are efforts on Wikipedia to translate articles, but limiting these translation efforts are the dynamic nature of articles, lack of multilingual editors, and editor interests. Multilingual Wikipedia editors are a highly scarce resource. We found 178,831 common usernames (registered editors) between English and German Wikipedias, which are 1.78% and 14.16% of the usernames of the respective Wikipedias. There are more anonymous editors, but the low total percentage of editors highlights the problem that few multilingual editors exist to transfer knowledge between languages.

In this poster, we present similarity and activity measures of multilingual articles. Our assumption is that Wikipedia articles may be written differently in different languages, but the knowledge or semantics they cover are similar based on the words used. There is evidence suggesting there is a consistency in the terminologies used across languages to convey ideas, knowledge, and facts [5, 3]. Furthermore, bilingual lexicons used by bilingual speakers can converge [3], which allows us to evaluate measures using machine translated versions of articles.

We provide an initial investigation into summarizing the vast Wikipedia corpus across different languages. Our aim is to develop a framework for recommending and verifying articles written in multiple languages by multilingual editors. The recommendation models allow us to suggest articles that can benefit from the different language skill levels of editors. This is important to expand and enrich content in smaller Wikipedias, where there may be editors learning or know those languages, but are unsure of what they can contribute. The verification models provide a unique way of determining knowledge by consensus across many languages. For example, parts of an article written by different editors that have similar knowledge representation across many languages are more likely to be trusted. This is important in document collections beyond Wikipedia, such as those produced by intergovernmental organizations, like the United Nations, to communicate between countries.

Our contributions include (1) novel similarity and activity measures, and (2) evaluation of these measures on over 620,000 Wikipedia articles written in two languages. From these results, we briefly discuss a framework for article recommendation and verification for multilingual editors.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
Copyright 2013 ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.250782.

Table 1: Basic statistics of data sets

Data set	All articles	Articles of interest	All revisions	Unique usernames	Unique IP addresses
English (en)	12,389,353	3,736,370	305,821,091	10,025,768	55,042,902
German (de)	2,826,811	1,235,009	65,732,032	1,262,688	12,511,832
Common	-	624,016 (en 16.70%, de 50.53%)	-	178,831 (en 1.78%, de 14.16%)	713,427 (en 1.30%, de 5.70%)

2. RELATED WORK

Sentence similarity is a related research area that looks at identifying highly similar sentences between two documents in different languages. For Wikipedia, similar sentences can be found by machine translation and the cross language link structure in articles [1]. These similar sentences can improve statistical machine translators by providing parallel sentences [10], and identify gaps of knowledge for Wikipedia editors [7].

These gaps of knowledge are also apparent in structural parts of Wikipedia such as templates and information boxes. For example, information boxes from different languages can be aligned, where missing facts from one language can be filled with information from other languages [2].

The semantic convergence of articles across languages can be determined by evaluating the similarity of sequential revisions of Wikipedia articles. When the similarity of content between revisions falls below a threshold, continuing edits do not change the meaning of articles [12].

Across many Wikipedia languages, similarities and differences in articles is much greater than previously assumed, as shown by Omnipedia [4]. Omnipedia shows information that is unique to each language, and other information shared across languages. A user study of Omnipedia shows how people interact with multilingual information and highlights the vast knowledge gaps in articles across many languages.

3. WIKIPEDIA DATA SETS

We use the Wikipedia (complete edit history) data dump of 1 June 2012 for the English (en) Wikipedia, and 3 June 2012 for the German (de) Wikipedia[1]. For both languages, we extract articles in Wikipedia's namespace 0, which contains only encyclopedic articles. We further remove articles that do not conform to Wikipedia markup and do not have much encyclopedic content, such as article stubs and disambiguation articles. We remove all Wikipedia markup from the article content for translation.

Table 1 shows some statistics of the resulting data set. A revision captures information about a change made to a Wikipedia article, such as editor, time, and content. Common articles are articles in both languages that have returning interlingual links. There are a small proportion of articles with no returning link from the other language. The common registered editors found in both Wikipedias are 1.78% of the English editors, and 14.16% of the German editors. The count of the usernames is of registered editors who have made an edit on Wikipedia that is recorded in the data dump.

4. MOSES MACHINE TRANSLATION

The Moses translator [8] is an offline free and open source statistical machine translator. We build a baseline Moses system, use the open source IRST LM Toolkit[2], and the Europarl Parallel Corpus and News Commentary corpus suggested for the Moses translator. Our training data set is the Europarl German-English corpus of 1,920,209 parallel sentences, and our tuning set is the News Commentary news-test2008 corpus of 2,051 parallel (occurring in both languages) sentences.

To evaluate performance of Moses, we translate the testing set news-test2011 from German to English with Moses, Google Translate, and Bing Translator. We use the BLEU (Bilingual Evaluation Understudy) score [9], a common evaluation method for machine translation system [6]. The BLEU scores are 0.1718, 0.1336, and 0.1493, for Moses, Google, and Bing, respectively. The Moses Translator is suitable for our research as it shows higher performance than two commercial statistical machine translators.

For comparison, the BLEU score for French are 0.23 for the baseline Moses algorithm, and around 0.30 for the best BLEU score at the Workshops on Statistical Machine Translation [6]. The lower scores observed for German may be from features of the German language, such as compound words, where the training set may not have a complete set of examples. The BLEU scores show the baseline Moses translator is a suitable alternative to commercial statistical machine translators.

5. MULTILINGUAL SIMILARITY

The multilingual similarity of two articles in two different languages is determined by analyzing the words in the articles and its translated equivalents. We assume an article written in two languages shares ideas, knowledge, and facts that transcends both cultures, despite different cultural contexts. We only look at the current revision of a Wikipedia article in both languages. We calculate the similarity on a *pivot* language. For example, using English as a pivot language, we calculate the similarity of the English article and the English translation of the German article.

For each current revision, we remove stopwords and punctuation relevant to each language, and clean other irrelevant tokens created by Moses. We use a TF-IDF representation of articles to determine the frequency and importance of occurring words within these articles. The Gensim[3] topic modeling toolkit is used to calculate the TF-IDF of words from the English and from the German collection of articles.

We use similarity measures common in information retrieval research: the Jaccard index, Dice's coefficient, and the Cosine similarity. These measures are also used in related work for monolingual and bilingual text. Due to space limitations, we present only results for the Cosine similarity with a TF-IDF modification. We cannot use the BLEU score as a similarity measure because the articles in English

[1] `http://dumps.wikimedia.org/`

[2] `http://hlt.fbk.eu/en/irstlm`

[3] `http://radimrehurek.com/gensim/`

and German are not sentence aligned. Finding these aligned sentences is research beyond the scope of this poster.

For each article, we have

$$sim_{l,i} = \frac{X_{l,i} \cdot Y_{l,i}}{\|X_{l,i}\| \|Y_{l,i}\|}, \tag{1}$$

where l is the pivot language such as English, i is the article identifier, $X_{l,i}$ is the TF-IDF vector of the English article, and $Y_{l,i}$ is the TF-IDF vector of the German article translated to English. The values of $sim_{l,i}$ are between 0 and 1, where 1 indicates high similarity.

6. ARTICLE ACTIVITY

We measure the activity of a Wikipedia article by analyzing the frequency and size of revisions. For each revision, we extract the time of the modification and the size of content of the article. We look at revisions of Wikipedia articles over their entire lifetime and in the most recent year covered by the data (2012) to compare past and recent activity. Due to space limitations, we present only one activity measure.

We measure activity by looking at the convergence of article revisions to a stable state. This assumes each article has a knowledge saturation point, where additional revisions do not change the meaning of the article. We model the growth of articles as logarithmic, because articles often have high frequency of changes and increasing article size early in their existence, but gradually decreasing in frequency of changes and smaller increases in article size. We use the absolute time in seconds since epoch and absolute size of the article content in bytes for each revision.

To determine growth, we take the log of the seconds and bytes, and perform a simple linear regression. We use the gradient as our measure of activity, where a small gradient indicates low activity. For each article in each language, we have

$$act_i = \frac{2}{\pi} \arctan \left| \frac{n \sum_j t_{i,j} s_{i,j} - \sum_j t_{i,j} \sum_j s_{i,j}}{n \sum_j (t_{i,j})^2 - \left(\sum_j t_{i,j} \right)^2} \right| \tag{2}$$

where $t_{i,j}$ is the time of revision j in seconds and $s_{i,j}$ is the size of revision j in bytes, and n is the total number of revisions for article i. The values of act_i are between 0 and 1, where 1 indicates high activity.

7. EVALUATION AND DISCUSSION

The measures presented provide a summary of knowledge coverage across languages and the relative activity of those articles within one language. We discuss the results of the measures individually, and only present plots of these two measures against each other.

For the Cosine TF-IDF similarity measure, our results show over 90% of articles have less than 0.75 similarity both when using either English or German as pivot languages. The Jaccard measure shows a much lower similarity for 90% of articles, at approximately 0.25 for English, and 0.2 for German. The Dice and Cosine measures show almost identical distribution of articles, where 90% of articles have a similarity less than 0.44 for English and 0.38 for German.

The Cosine TF-IDF results suggest most articles written in English and German share many common important words in both languages in describing the same topic. However, when we do not adjust for the importance of words

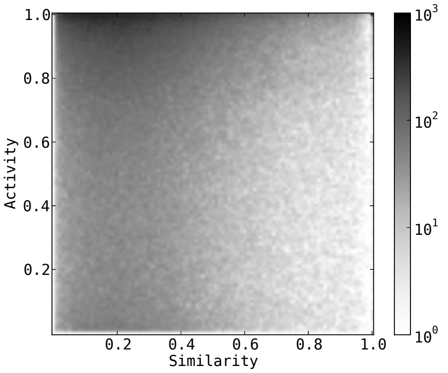

Figure 1: Heatmap of the number of articles with similar activity and similarity values in logarithmic scale. English as pivot language. 2012 revisions.

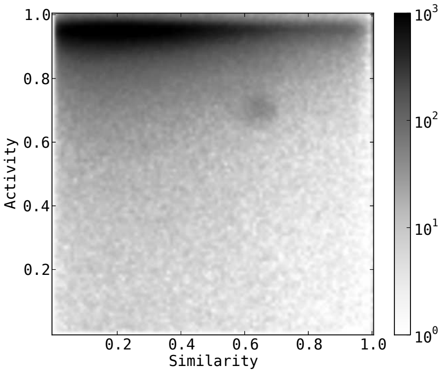

Figure 2: English as pivot language. All revisions.

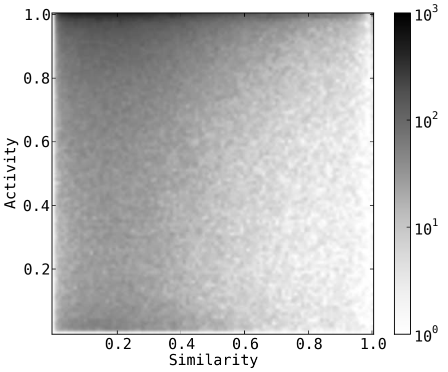

Figure 3: German as pivot language. 2012 revisions.

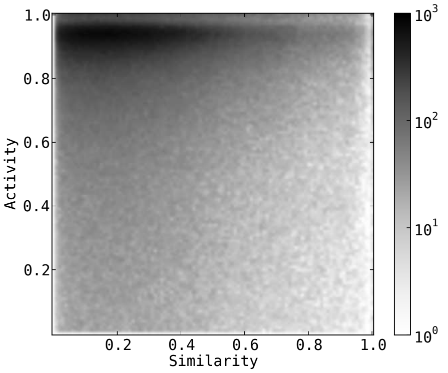

Figure 4: German as pivot language. All revisions.

in articles using TF-IDF, the similarity drops significantly. The higher similarity observed for the English pivot language suggests German articles may be using their English article equivalents as a source of knowledge. The limitation of measuring similarity with translated text is the quality of the machine translator and the training data set.

For the Gradient activity measure from Equation 2 calculated over all revisions, our results show 90% of articles have an activity level over 0.8 for English, and over 0.5 for German. For the recent 2012 revisions, over 37% of articles for English and over 55% for German have an activity of 0.0; and less than 32% of active (non-zero activity) articles have an activity level over 0.8 for English and less than 24% for German.

These results suggest that over the lifetime of most articles, activity is seen in some form. For recent activity, we see more than half of the German version of articles has no activity. Of the articles with non-zero activity level, we observe higher activity levels in the English version compared to the German version. This may be explained by the higher number of editors working in English.

Plotting these two measures against each other shows the distribution of activity on articles across articles with different levels of knowledge coverage. We compare recent and all revisions, and different pivot languages in Figures 1 to 4. The figures show a heatmap of the number of articles based on their similarity and activity values. For the activity measure over all revisions, we see the English version of the articles generally have higher levels of activity. For the recent 2012 revisions, we see the English articles have much less activity than the German articles. The similarity shows a general grouping towards lower similarity when viewed in both English and German pivot languages.

These results show the distribution of articles based on the proposed measures. We observe some forms of clusters, which we will further explore in future work. In particular, there is a clear distinction of articles with high and low activity in the recent 2012 revisions. The four corners of each figure present options for recommendation to editors based on language skill. For example, low activity and high similarity articles (lower right corner) suggest filling in knowledge gaps, which may be suitable for editors learning a new language. The high similarity articles allow knowledge verification models to be developed from the content as these articles show much shared knowledge. These verification models can be used to improve trust of content from readers, and low similarity multilingual articles can be written to be consistent across languages.

8. CONCLUSION

We have presented similarity and activity measures for summarizing the knowledge coverage of articles over different languages, and the level of activity on each article, respectively. We apply these measures to Wikipedia articles written in both English and German. We translate articles of both languages into the other to compare differences in the similarity measure when viewed using English or German as the pivot language. Our results show high knowledge coverage when we consider words important to the article. Our activity measure summarizes the level of activity of articles in each language. Our results show English versions of articles have a generally higher activity level for both recent and all revisions. Our plots of activity against similarity show the distribution of articles, where we briefly discuss future clustering, recommendation, and verification work.

In future work, we aim to complete cluster analysis of these measures, and use the translated articles to improve document clustering [13]. We also aim to develop prototype recommendation and verification systems, evaluated by user studies. We aim to add more languages and use each as a pivot language for analysis. We plan to explore direct word translation methods as translating text with document context is a computationally intensive task.

9. REFERENCES

[1] S. Adafre and M. de Rijke. Finding similar sentences across multiple languages in wikipedia. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[2] E. Adar, M. Skinner, and D. S. Weld. Information arbitrage across multi-lingual Wikipedia. In *2nd ACM International Conference on Web Search and Data Mining*, 2009.

[3] E. Ameel, B. Malt, G. Storms, and F. Van Assche. Semantic convergence in the bilingual lexicon. *Journal of memory and language*, 60:270–290, 2009.

[4] P. Bao, B. Hecht, S. Carton, M. Quaderi, M. Horn, and D. Gergle. Omnipedia: Bridging the Wikipedia Language Gap. In *Proc. CHI 2012*, 2012.

[5] L. Boroditsky. How language shapes thought. *Scientific American*, 304(2):62–65, 2011.

[6] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. Findings of the 2011 Workshop on Statistical Machine Translation. In *6th Workshop on Statistical Machine Translation*, 2011.

[7] E. Filatova. Multilingual wikipedia, summarization, and information trustworthiness. In *SIGIR Workshop on Information Access in a Multilingual World*, 2009.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In *45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 2007.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics*, 2002.

[10] J. R. Smith, C. Quirk, and K. Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. In *2010 Annual Conference of the North American Chapter of the ACL*, 2010.

[11] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: slowing growth of Wikipedia. In *5th International Symposium on Wikis and Open Collaboration*, 2009.

[12] C. Thomas and A. P. Sheth. Semantic Convergence of Wikipedia Articles. In *IEEE/WIC/ACM International Conference on Web Intelligence*, 2007.

[13] X. Wang, B. Qian, and I. Davidson. Improving document clustering using automated machine translation. In *21st ACM International Conference on Information and Knowledge Management (CIKM)*, 2012.