Multilingual Wikipedia, Summarization, and Information Trustworthiness

Elena Filatova
Fordham University
Department of Computer and Information Sciences
filatova@cis.fordham.edu

ABSTRACT

Wikipedia is used as a corpus for a variety of text processing applications. It is especially popular for information selection tasks, such as summarization feature identification, answer generation/verification, etc. Many Wikipedia entries (about people, events, locations, etc.) have descriptions in several languages. Often Wikipedia entry descriptions created in different languages exhibit differences in length and content. In this paper we show that the pattern of information overlap across the descriptions written in different languages for the same Wikipedia entry fits well the *pyramid* summary framework, i.e., some information facts are covered in the Wikipedia entry descriptions in many languages, while others are covered in a handful number of descriptions. This phenomenon leads to a natural summarization algorithm which we present in this paper. According to our evaluation, the generated summaries have a high level of user satisfaction. Moreover, the discovered pyramid structure of Wikipedia entry descriptions can be used for Wikipedia information trustworthiness verification.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Measurement, Experimentation, Human Factors

Keywords

Wikipedia, summarization, multilinguality

1. INTRODUCTION

Wikipedia[1,2] provides descriptions of people, events, locations, etc. in many languages. Despite the recent discussion of the Wikipedia descriptions trustworthiness or lack of thereof [9], Wikipedia is widely used in information retrieval (IR) and natural language processing (NLP). The question arises: what can be done to increase the trustworthiness of the information extracted from Wikipedia. We believe, Wikipedia itself has resources to increase its trustworthiness.

Most Wikipedia entries have descriptions in different languages. These descriptions are *not* translations of a Wikipedia entry description from one language into other languages. Rather, Wikipedia entry descriptions in different languages are independently created by different users. The length of the entry descriptions about the same Wikipedia entry varies greatly from language to language. Obviously, texts of different length cannot contain the same amount of information about an entry.

In this paper we compare descriptions of Wikipedia entries written in different languages and investigate the information overlap pattern in multilingual Wikipedia. We show that information overlap in entry descriptions written in different languages corresponds well to the pyramid summarization model [16, 12]. This result helps the understanding of the combined value of the multilingual Wikipedia entry descriptions. On the one hand, multilingual Wikipedia provides a natural summarization mechanism. On the other hand, to get a complete picture about a Wikipedia entry, descriptions in all languages should be combined. Finally, this pyramid structure can be used for information trustworthiness verification.

The rest of the paper is structured as follows. In Section 2 we describe related work. In Section 3 we provide a motivation example for our research. In Section 4 we describe our corpus, the summarization-based experiments we ran to analyze multilingual Wikipedia information overlap. In Section 5 we discuss the results of our experiments. In Section 6 we outline the avenues for future research.

2. RELATED WORK

Multilingual aspect of Wikipedia is used for a variety of text processing tasks. Adafre *et al.* [8] analyze the possibility of constructing an English-Dutch parallel corpus by suggesting two ways of looking for similar sentences in Wikipedia pages (using matching translations and hyperlinks). Richman *et al.* [13] utilize multilingual characteristics of Wikipedia to annotate a large corpus of text with Named Entity tags. Multilingual Wikipedia is used to facilitate cross-language IR [14] and to perform cross-lingual QA [6].

The described applications do not question the trustworthiness of the information presented in Wikipedia. In a separate line of research, approaches are developed to rate the trustworthiness of Wikipedia information. These approaches, however, deal with the text only one language.

Wikipedia content trustworthiness can be estimated using a combination of the amount of the content revision and the author reputation performing this revision [2]. Another way to use edit history to estimate information trustworthiness is to use dynamic Bayesian network trust model that utilized rich revision information in Wikipedia. [17]. Wikipedia trustworthiness can be also estimated using an additional tab (*Trust tab*) [11].

[1] `http://en.wikipedia.org/wiki/Wikipedia`

[2] Wikipedia is changing constantly. All the quotes and examples from Wikipedia presented and analyzed in this paper were collected on February 10, 2009, between 14:00 and 21:00 PST.

The research closest to ours was recently described in Adar *et al.* [1] where the main goal is to use self-supervised learning to align or/and create new Wikipedia infoboxes across four languages (English, Spanish, French, German). Wikipedia infoboxes contain a small number of facts about Wikipedia entries in a semi-structured format. In our work, we deal with plain text and disregard any structured data such a infoboxes, tables, etc. It must be noted, that the conclusions that are reached in parallel for structured Wikipedia information by Adar *et al.* and for unstructured Wikipedia information by us are very similar. These conclusions stress the fact that the most trusted information is repeated in the Wikipedia entry descriptions in different languages. At the same time, no single entry descriptions can be considered as the complete source of information about a Wikipedia entry.

3. INFORMATION OVERLAP

Currently, Wikipedia has entry descriptions in more than 200 languages. The language with the largest number of entry descriptions is English [8, 5] but the size of non-English Wikipedia is growing fast and represents a rich corpus.[3]

Most existing NLP applications that use Wikipedia as the training corpus or information source assume that Wikipedia entry descriptions in all languages are a reliable source of information. However, according to our observations, Wikipedia descriptions about the same entry (person, location, event, etc.) in different languages frequently cover different sets of facts. According to the Wikipedia analysis [7], there are two major sources of differences in the descriptions of the same Wikipedia entry written in different languages:

- the amount of information covered by a Wikipedia entry description;[4]
- the choice of information covered by a Wikipedia entry description.

In this paper we analyze the information overlap in Wikipedia entry descriptions written in several languages.

For example, baseball is popular in the USA, Latin America, and Japan but not in Europe or Africa. Wikipedia has descriptions of *Babe Ruth* in 18 languages: the longest and most detailed descriptions are in English, Spanish and Japanese. The description of *Babe Ruth* in Finnish has five and in Swedish - four sentences. These short entry descriptions list several general biographical facts: dates of birth, death; the fact that he was a baseball player. It is likely, that the facts from the Swedish and Finnish entry descriptions about *Babe Ruth* will be listed in a summary of the English language Wikipedia entry description of him.

Currently, information overlap is successfully used for summarization evaluation. The state-of-the-art (automatic and manual) summarization evaluation approaches compare the target summary against several model summaries. Such models are typically created manually. The more model summaries contain a specific piece of information - the greater value it gets in the target summary [10, 16, 12].

4. CORPUS ANALYSIS EXPERIMENT

In this paper, we investigate how the information overlap in multilingual Wikipedia can be used to create summaries of entry descriptions.

[3] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[4] In this work, the length of a Wikipedia entry description is measured in sentences used in the text description of a Wikipedia entry.

4.1 Data Set

For our experiment, we used the list of people created for the Task 5 of DUC 2004: biography generation task (48 people).[5] This set is small enough to be analyzed manually in detail; at the same time, it is widely used for summarization experiments. In the future, we plan to compare the performance of our summarization approach against another summarization system developed for biography generation that utilizes Wikipedia document structure [4].

We downloaded from Wikipedia all the entry descriptions in all the languages corresponding to each person from the DUC 2004 list. We used Wikitext, the text that is used by Wikipedia authors and editors. Wikitext can be obtained through Wikipedia dumps.[6] We removed from the wikitext all the markup tags and tabular information (e.g., infoboxes and tables) and kept only plain text. There is no commonly accepted standard wikitext language, thus our final text had a certain amount of noise which, however, as discussed in Section 5, did not affect our experimental results.

For each Wikipedia entry (i.e., DUC 2004 person) we downloaded corresponding entry descriptions in all the languages, including Esperanto, Latin, etc. We used the name of a person to find the description of this person in English and then we followed the links from the left side panel of the Wikipedia page template to get the entry descriptions in other languages. To facilitate the comparison of entry descriptions written in different languages we used the Google machine translation tool[7] to translate the downloaded entry descriptions into English. The number of languages covered currently by the Google translation system (41) is less than the number of languages used in Wikipedia (265). However, the language distribution in the collected corpus corresponds well the language distribution in Wikipedia and the collected Wikipedia subset can be considered a representative sample [7].

Five people from the DUC 2004 set had only English Wikipedia entry descriptions: *Paul Coverdell*, *Susan McDougal*, *Henry Lyons*, *Jerri Nielsen*, *Willie Brown*. Thus, they were excluded from the analysis. The person whose Wikipedia entry had descriptions in most languages (86) was *Kofi Annan*. On average, a Wikipedia entry for a DUC 2004 person had descriptions in 25.35 languages. The description in English was not always the longest description: in 17 cases the longest description of a Wikipedia entry for a DUC 2004 person was in a language other than English.

4.2 Data Processing Tools

After the Wikipedia entry descriptions for the DUC 2004 set were collected and translated, we divided these descriptions into sentences using the LingPipe sentence chunker [3]. For each DUC 2004 person we compared a description of this person in English against the descriptions of this person in other languages that were handled by the Google translation system. We counted descriptions in how many languages had sentences corresponding to the sentences in the description in English. To identify matching sentences we used the LingPipe string matching tool based on TF/IDF distance which "*is based on vector similarity (using the cosine measure of angular similarity) over dampened and discriminatively weighted term frequencies. [...] two strings are*

[5] http://duc.nist.gov/duc2004/tasks.html/

[6] http://download.wikimedia.org/

[7] http://translate.google.com/

	Algorithm
1	Submit the person's name to Wikipedia
2	Get Wikipedia entry descriptions for this person in all possible languages
3	Remove non-plain text information from the descriptions
4	For all the languages handled by the Google MT, translate entry descriptions into English
5	Break English texts into sentences
6	Use a similarity measure to identify what English sentences have counterparts in entry descriptions in other languages
7	Rank all the sentence from the English document according to the number of languages that have similar sentences
8	If several sentences are placed on the same level, list these sentence in the order they appear in the Wikipedia entry description in English
9	Use the top three levels from the above ranking

Table 1: Algorithm outline.

more similar if they contain many of the same tokens with the same relative number of occurrences of each. Tokens are weighted more heavily if they occur in few documents" [3]. For our experiment, each sentence was treated as a separate document. The IDF value was computed based on the two entity descriptions under consideration (one - in English, the other one - translation into English). We used three similarity thresholds: 0.5, 0.35, 0.2.

4.3 What was Measured

To evaluate how much information is repeated in the descriptions of the same person in different languages we measured similarity of the person's description in English and in other languages. As each sentence was treated as a separate document, the number of tokens (words) for comparison was rather small. Thus, for the 0.5 similarity threshold, the sentences marked as similar were almost identical. Using the 0.35 and 0.2 thresholds allowed to search for non-identical sentences that still had a substantial word overlap.

Our hypothesis is that those facts (sentences) that are mentioned in the descriptions of a person in different languages fit well the pyramid summarization model. For example, if we are to summarize a description of a person from the English Wikipedia: first, we should add to the summary those sentences that have their counterparts in the most number of descriptions of this person in the languages other than English. Sentences added on this step correspond to the top level of the pyramid. If the length of the summary is not exhausted then, on the next step, we add to the summery those sentence that appear in the next most number of languages, and so on. Thus, we can place sentences on different levels of the pyramid, with the top level being populated by the sentences that appear in the most languages and the bottom level having sentences that appear in the least number of languages. For our experiments we used sentences from the top three levels of this pyramid. All the sentences added to the summary should appear in at least two languages other than English. Table 1 has a schematic outline of the described algorithm.

4.4 Example and Experiment Discussion

Table 2 presents three-level summaries for the English Wikipedia description of *Gene Autry*. Wikipedia has descriptions of *Gene Autry* in 11 languages: in English and in ten other languages each of which can be translated into English by the Google Translation system.

Using similarity of 0.5 we get one sentence from the English description that has counterparts in at least two other languages (here, in three other languages). This sentence's ID is 0: it is the entry description introductory sentence.

Using similarity 0.35 we get a summary consisting of six sentences. The sentence on the top level is the same sentence that was listed in the previous summary. However, having a more permitting similarity threshold, this sentence was mapped to similar sentences in 7 languages, rather than in 3. Next level consists of those sentences from the English description that were mapped to sentences in three other languages. Sentences on the third level were mapped to sentences from two other languages. It is interesting to notice that the sentences included in the summary are coming from different parts of the document that has 88 sentences.

The summary created using the 0.2 threshold contains the introductory sentence as well as sentences not included in the summaries for the 0.5 and 0.35 similarity threshold.

Despite the fact that for our experiment we chose the set of people used for the DUC 2004 biography generation task, we could not use the DUC 2004 model summaries for our evaluation. These models were created using the DUC 2004 corpus, while in our experiments we used a subset of multilingual Wikipedia. Moreover, Wikipedia entry descriptions about the DUC 2004 people had dramatic updates since 2004. For example, *Jörg Haider* died of injuries from a car crash on October 11, 2008 and this information is included into our three-level summaries.

Due to space constraints, in this paper we report only the results obtained using similarity threshold of 0.35. Also, in the experiment described in this paper we analyze only those sentences from the English text that appear in at least two other languages, with the exception for *Louis Freeh*, for whom only one language was handled by the Google Translation system. Thus, the summary for the English entry description about *Louis Freeh* has only one level which has all the sentences from the English entry description that have their counterparts in the only available translation. Thus, using the DUC 2004 set we created:

- **one**-level summaries for 5 people;
- **two**-level summaries for 3 people;
- **three**-level summaries for 35 people.

The length of the created summaries is measured in sentences. Table 3 presents information about the average and maximal length of summaries for all three levels combined and for each level separately. The summaries that do not have Level 2 and/or Level 3 are excluded from the corresponding average and maximum value computation. According to the presented data, on average, the output three-level summaries are rather short, however, some summaries can be quite long. We believe that such a difference between the average and the maximal length is due to:

1. the length variation of the English Wikipedia entry descriptions;
2. the number variation of descriptions (languages) for each a person and the lengths of these descriptions.

To evaluate the output three-level summaries we used Amazon Mechanical Turk as a source of human subjects who can reliably evaluate certain NLP tasks [15]. For each of the 43 outputs (for 43 people from the DUC 2004 set) we

#	Lang.	Sent. ID	Text
Similarity 0.5			
1	3	0	Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as The Singing Cowboy on the radio, in movies and on television.
Similarity 0.35			
1	7	0	Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as "The Singing Cowboy" on the radio, in movies and on television.
2	3	1	Autry, the grandson of a Methodist preacher, was born near Tioga, Texas.
		13	His first hit was in 1932 with "That Silver-Haired Daddy of Mine," a duet with fellow railroad man, Jimmy Long.
3	2	3	After leaving high school in 1925, Autry worked as a telegrapher for the St. Louis-San Francisco Railway.
		14	Autry also sang the classic Ray Whitley hit "Back in the Saddle Again," as well as many Christmas songs including "Santa Claus Is Coming to Town," his own composition "Here Comes Santa Claus," "Frosty the Snowman," and arguably his biggest hit "Rudolph the Red-Nosed Reindeer."
		72	Gene Autry died of lymphoma at age 91 at his home in Studio City, California and is interred in the Forest Lawn, Hollywood Hills Cemetery in Los Angeles, California.
Similarity 0.2			
1	7	0	Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as "The Singing Cowboy" on the radio, in movies and on television.
2	6	1	Autry, the grandson of a Methodist preacher, was born near Tioga, Texas.
		73	His death on October 2, 1998 came nearly three months after the death of another celebrated cowboy of the silver screen, radio, and TV, Roy Rogers.
3	5	21	From 1940 to 1956, Autry had a huge hit with a weekly radio show on CBS, "Gene Autry's Melody Ranch." His horse, Champion, also had a radio-TV series "The Adventures of Champion."
		72	Gene Autry died of lymphoma at age 91 at his home in Studio City, California and is interred in the Forest Lawn, Hollywood Hills Cemetery in Los Angeles, California.

Table 2: Three-level summaries for Gene Autry (*#*: the summary level; *Lang.*: number of languages that contain a similar sentence; *Sent. ID*: the position of the sentence in the English description; *Text*: the sentence itself).

	Three-level summary	Level one	Level two	Level three
Avg	3.74	1.02	1.58	1.63
Max	9	2	6	7

Table 3: Summaries length: average and maximal.

recruited five human annotators. The annotators were provided with: the name of the person; link to the Wikipedia entry description about this person in English; three-level summary of this Wikipedia entry description. We asked our human annotators to answer the following questions:

- Do you agree that the sentences listed on Level 1 are a good summary of the Wikipedia entry description about *Person* (assume, the number of sentences in the summary cannot exceed the number of sentences listed on Level 1)?
- Assume that the summary of the Wikipedia entry description about *Person* can have as many sentences as listed on Level 1 and Level 2 combined. Do you agree that the sentences listed on Level 1 and Level 2 are a good summary?
- Assume that the summary of the Wikipedia entry description about *Person* can have as many sentences as listed on Level 1, Level 2, and Level 3 combined. Do you agree that the sentences listed on Level 1, Level 2, and Level 3 are a good summary?

If the summary did not have Level 2 and/or Level 3 sentences, the annotator was asked to skip answering the corresponding questions.

5. RESULTS

Table 4 summarizes the results of the three-level summaries evaluation. The **Goodness** measure shows how many (out of five) annotators agreed that the summary for a particular level, given the length constraint, was good. The numbers in the table show the number of summaries that were considered good for each level according to a particular level of goodness. As it is mentioned in Section 4.4 not all summaries have Levels 2 and 3 filled in; the *Number of summaries* column in Table 4 has this information.

According to Table 4, no summary on Level 1 was uniformly considered bad. One summary was considered bad by four out of five annotators. This was the summary for *Paul Wellstone* with Level 1 consisting only of one sentence. We analyzed this sentence and discovered that it was incorrectly truncated due to our sentence chunker error.

> [Paul David Wellstone (July 21, 1944 - October 25, 2002) was a two-term U.S.] [Senator from the U.S. state of Minnesota and member of the Democratic-Farmer-Labor Party, which is affiliated with the national Democratic Party.]

This sentence was broken into two sentences (identified above by the square brackets), and only the first portion of the sentence was added to the Level 1 summary. Despite the fact that this portion contains important biographical information, it cannot be used as a stand-alone sentence. According to our analysis, three out of seven summaries that were judged as bad by three out of five annotators had exactly the same problem of incorrect sentence segmentation forcing only portions of sentences to be added to the summaries.

In addition to asking annotators to judge the quality of the created summaries we welcomed our annotators to leave comments about the summaries they read. These comments can be divided into two groups. Several annotators noticed text preprocessing errors (e.g., leftovers from the Wikitext XML tagging): however, this did not affect their judgement of the summary quality: all the summaries containing XML tags were marked as good. The other set of observations

	Goodness						Number of
Levels	**5**	**4**	**3**	**2**	**1**	**0**	summaries
1	28	4	3	7	1	0	43
1,2	12	3	12	4	5	2	38
1,2,3	5	6	14	8	1	1	35

Table 4: Evaluation results (using Mechanical Turk).

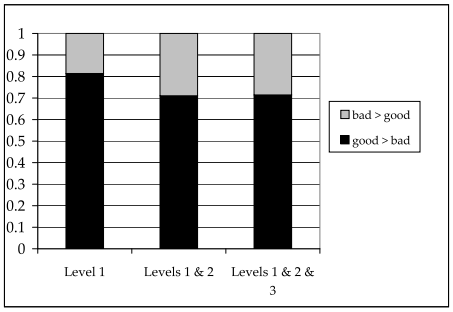

Figure 1: Combined results.

concerned the type of facts included in the summaries. For example, it was pointed out that the sentences from the summary about *Abdullah Öcalan* did not have **enough** information about his political activities and thus, the created summaries were judged as bad. Annotators suggested that information about professional life of politicians would be more appropriate than the information about their marriages. However, sentences containing information about private life were mostly considered relevant and judged as good additions to summaries.

Figure 1 shows the combined numbers for Table 4. For each level we grouped all the numbers in two categories: those numbers where the majority of the annotators agreed that the summary was good and those numbers where the majority of the annotators decided that the summary was bad. As not all the summaries had sentences from all three levels, Figure 1 has ratios rather than the absolute numbers listed in Table 4. This figure shows that overall the quality of the created summaries was quite high. In more than 80% of cases our annotators were happy with the summaries consisting of the Level 1 sentences, and in more that 70% of cases our annotators were happy with the summaries consisting of the sentences combined from Levels 1 and 2 and Levels 1, 2, and 3. To conclude this section, we showed that information overlap in multilingual Wikipedia can be used for placing information facts into a pyramid structure.

6. FUTURE WORK

While the main focus of the current paper is information overlap, in the future, we are interested in studying information asymmetries in multilingual Wikipedia.

We are interested in investigating how Wikipedia multilinguality can be used for opinion, contradiction and new information detection. An important observation concerning the example presented in Section 3 is that irrespectively of the length of the descriptions of a person in different languages, none of these descriptions have any facts that contradict the facts in the descriptions of this Wikipedia entry in other languages. Rather, the discussed entry descriptions in different languages contain a subset of facts that appear in many languages plus, maybe, additional information. This allowed us to formulate and test the hypothesis that a set of Wikipedia entry descriptions about the same entry fits well the pyramid summarization model. However, there are Wikipedia entries attitude to which is different in communities speaking in different languages. In these cases, we believe, it will be more interesting to draw the readers attention not to the facts that are repeated in the entry descriptions in many languages but rather, highlight differences among these descriptions.

7. REFERENCES

[1] E. Adar, M. Skinner, and D. Weld. Information arbitrage in multi-lingual Wikipedia. In *WSDM*, 2009.

[2] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *WikiSym*, 2008.

[3] Alias-i. LingPipe 3.7.0. (accessed January 19, 2009), 2009. `http://alias-i.com/lingpipe`.

[4] F. Baidsy, J. Hirschberg, and E. Filatova. An unsupervised approach to biography production using Wikipedia. In *ACL*, 2008.

[5] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[6] S. Ferrández, A. Toral, O. Ferrández, A. Ferrández, and R. Munoz. Applying Wikipedia's multilingual knowledge to cross-lingual Question Answering. *LNCS: Natural Language Processing and Information Systems*, 4592:352–363, 2007.

[7] E. Filatova. Directions for exploiting asymmetries in multilingual Wikipedia. In *CLIAW3*, 2009.

[8] S. Fissaha Adafre and M. de Rijke. Finding similar sentences across multiple languages in Wikipedia. In *Workshop on New Text – Wikis and blogs and other dynamic text sources*, 2006.

[9] A. Keen. *The Cult of the Amateur: How Today's Internet is Killing Our Culture*. Bantam Books, 2007.

[10] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *HLT-NAACL*, 2003.

[11] D. McGuinness, H. Zeng, P. Pinheiro da Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigations into trust for collaborative information repositories: A Wikipedia case study. In *MTW*, 2006.

[12] A. Nenkova, R. Passonneau, and K. McKeown. The Pyramid method: Incorporating human content selection variation in summarization evaluation. *TSLP*, 4(2), 2007.

[13] A. Richman and P. Schone. Mining Wiki resources for multilingual named entity recognition. In *ACL*, 2008.

[14] P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány. Performing cross-language retrieval with Wikipedia. In *CLEF*, 2007.

[15] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *KDD*, 2008.

[16] S. Teufel and H. V. Halteren. Evaluating information content by factoid analysis: Human annotation and stability. In *ACL*, 2004.

[17] H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. McGuinness. Computing trust from revision history. In *PST*, 2006.