International Journal of Cooperative Information Systems
Vol. 21, No. 4 (2012) 1–21

DOI: 10.1142/S0218843013500019

INFORMATION OVERLAP IN MULTILINGUAL WIKIPEDIA AND SUMMARIZATION

ELENA FILATOVA

Computer and Information Sciences Department
Fordham University, 441 East Fordham Road
Bronx NY 10458, USA
filatova@cis.fordham.edu

Received 12 November 2011
Accepted 2 October 2012
Published

Wikipedia is used as a training corpus for many information selection tasks: summarization, question-answering, etc. The information presented in Wikipedia articles as well as the order in which this information is presented, is treated as the gold standard and is used for improving the quality of information selection systems. However, the Wikipedia articles corresponding to the same entry (person, location, event, etc.) written in different languages have substantial differences regarding what information is included in these articles. In this paper we analyze the regularities of information overlap among the articles about the same Wikipedia entry written in different languages: some information facts are covered in the Wikipedia articles in many languages, while others are covered only in a few languages. We introduce a hypothesis that the structure of this information overlap is similar to the information overlap structure (pyramid model) used in summarization evaluation, as well as the information overlap/repetition structure used to identify important information for multidocument summarization. We prove the correctness of our hypothesis by building a summarization system according to the presented information overlap hypothesis. This system summarizes English Wikipedia articles given the articles about the same Wikipedia entries written in other languages. To evaluate the quality of the created summaries, we use Amazon Mechanical Turk as the source of human subjects who can reliably judge the quality of the created text. We also compare the summaries generated according to the information overlap hypothesis against the lead line baseline which is considered to be the most reliable way to generate summaries of Wikipedia articles. The summarization experiment proves the correctness of the introduced multilingual Wikipedia information overlap hypothesis.

Keywords: Summarization; crowdsourcing; information overlap; multilingual wikipedia.

1. Introduction

Many information selection applications utilize multiple descriptions of the same object and often rely on information repetition in these multiple descriptions. Summarization applications use multiple descriptions of the input entry (person, event, etc.) to extract the most important information about the input entry. Answer

validation applications use the information extracted from different documents to identify which of the suggested answers is the correct one, etc.

In addition to processing multiple descriptions about the same entry, most information selection systems require the gold standard against which their output can be compared. Typically, the gold standard is created by human annotators: correct answers to questions, model summaries, etc. Creating the gold standard is often expensive and time-consuming as it requires a lot of hours of work conducted by trained human annotators.

Wikipedia[a] provides articles about people, events, locations, etc. in many languages. It is constantly monitored by Wikipedia editors to ensure that Wikipedia articles comply with the Wikipedia standards and is up-to-date. Many information selection systems utilize these Wikipedia features assuming that the information presented in Wikipedia is correct. Some systems use Wikipedia content as the source of reliable and up-to-date information, others utilize the structure of Wikipedia articles (see Sec. 2).

In this work, we raise an issue that certain Wikipedia peculiarities should be taken into consideration by the researchers that use Wikipedia as the source of the gold standard for information selection. Namely, most Wikipedia entries have articles in several languages. These articles are not translations of a Wikipedia article from one language into other languages. Rather, Wikipedia articles in different languages are independently created by different users. Thus, these documents are not *identical* but rather can be treated as *different descriptions* of the same entry. At the same time, despite all the differences, the Wikipedia articles about the same entry written in different languages obviously have a certain degree of information overlap. Some information facts are repeated in the articles in many languages while others are covered in the articles in only a handful of languages.

Given the fact that Wikipedia articles about the same entry written in different languages contain different sets of facts, the question arises whether or not Wikipedia can be used as the gold standard for information selection tasks. The issue of the trustworthiness of the Wikipedia information is currently studied for monolingual Wikipedias. Obviously, this issue becomes even more complex when multilingual Wikipedia is analyzed.

We compare descriptions of Wikipedia entries written in different languages and investigate their information overlap pattern. We introduce a hypothesis regarding the pattern of the information overlap in multilingual Wikipedia. According to our hypothesis: information overlap in Wikipedia articles about the same entry written in different languages corresponds well to the pyramid summarization evaluation model.[1,2]

To show the correctness of this hypothesis we create a summarization system that follows our information overlap hypothesis. As most extractive summarization systems work with sentences as the summary generation units, in our work

[a]http://en.wikipedia.org/wiki/Wikipedia

we use a sentence as a proxy for an information fact: the more articles in different languages contain a certain sentence (information fact) — the higher the priority for this sentence (information fact) to be included into the output summary. The summarization system that follows our multilingual Wikipedia information overlap hypothesis is close in nature to the multidocument summarization systems that rely on information repetition to rank information facts according to their importance.[3,4]

The summarization system created for our hypothesis testing combines both single- and multidocument summarization aspects: it summarizes a single document (English Wikipedia article about a certain entry) given multiple reference documents (Wikipedia articles about the same entry written in different languages). The evaluation experiment shows that the summarization system built according to our multilingual Wikipedia information overlap hypothesis has a high level of user satisfaction. We believe this result does not only prove the correctness of our hypothesis, but also helps the understanding of the combined value of multilingual Wikipedia entry descriptions. On the one hand, in order to get a complete picture about a Wikipedia entry, the articles in all languages should be combined.[b] On the other hand, the structure of the multilingual Wikipedia information overlap can be used for a variety of NLP tasks, including summarization, information trustworthiness estimation, etc.

The rest of the paper is structured as follows. In Sec. 2, we describe related work. In Sec. 3, we describe in detail our hypothesis about the nature of the information overlap in multilingual Wikipedia. In Sec. 4, we describe our corpus, the summarization-based experiments that we use to prove our hypothesis regarding the structure of the information overlap in multilingual Wikipedia, and discuss the results of this experiment. In Sec. 5, we outline the avenues for future research.

2. Related Work

Wikipedia has become a popular corpus for training and testing information selection systems.

Ahn *et al.*[7] use Wikipedia "both as a source of answers to factoid questions and as an importance model". Buscaldi *et al.*[8] use Wikipedia for answer verification; they also use "Wikipedia *categories* in order to determine a set of patterns that should fit with the expected answer". Ko *et al.*[9] use Wikipedia for measuring answer relevance and as a source of synonyms. In Refs. 7 and 8 monolingual (English and Spanish, respectively) Wikipedias are used. Ko *et al.*[9] use English, Chinese and Japanese Wikipedias. However, these three corpora are used independently and thus, one can say that Ref. 9 uses three versions of monolingual Wikipedia. In these papers, Wikipedia articles are treated as the gold standard: the systems do not question the correctness of the information presented in Wikipedia.

[b]The preliminary experiments and findings for this paper have been published in Refs. 5 and 6.

Baidsy *et al.*[10] use the structure of Wikipedia articles describing people to learn the order in which facts should be presented in a biography. Their results show that the summarization system that generates biographies following the structure of Wikipedia articles describing people "significantly outperforms all systems that participated in DUC 2004, according to the ROUGE-L metric, and is preferred by human subjects".

Many successful multidocument summarization systems heavily rely on information repetition to produce good quality summaries.[19–21] These systems use the idea that an important information fact that should be included into the final summary is likely to be repeated in many input documents. The idea of the parallel between information importance and its use in the summaries is the foundation of the pyramid-based summarization evaluation system.[2] The pyramid-based summarization evaluation approach relies on the model summaries created by humans: the more humans included a certain information fact or summary content unit (SCU) into the summary — the more important for a summarization system to include this information fact into its output.

Nelken *et al.*[11] use article revision history for collection and correction of lexical errors; for training sentence compression algorithms; and for discerning lexicalized models.

The multilingual aspect of Wikipedia is used for a variety of text processing tasks. The potential of Wikipedia as the source of parallel corpora is investigated in Ref. 12. The authors construct an English–Dutch parallel corpus and describe two ways of looking for similar sentences in Wikipedia pages (using matching translations and hyperlinks). Multilingual Wikipedia is used to annotate a large corpus of text with Named Entity tags in Ref. 13; to facilitate cross-language IR in Ref. 14, and to perform cross-lingual QA in Ref. 15.

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 monolingual Wikipedias.

One approach to compute the trustworthiness of Wikipedia information is to use the information from its edit history. Wikipedia content trustworthiness can be estimated by using a combination of the amount of the content revised and the author's reputation performing this revision Ref. 16. Another way to use edit history to estimate information trustworthiness is to treat Wikipedia article editing as a dynamic process and to use a dynamic Bayesian network trust model that utilizes rich revision information in Wikipedia.[17] Wikipedia trustworthiness can also be estimated by using an additional tab (*Trust tab*).[18]

The research closest to ours is presented in Ref. 22, where the main goal is to use self-supervised learning to align and/or 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, however, we deal with plain text and disregard any structured data such

a infoboxes, tables, etc. Although, we and Adar *et al.* analyze different types of Wikipedia information, we arrive to similar conclusions: the most trusted information is repeated in the Wikipedia articles in different languages. At the same time, no single article can be considered as the complete source of information about a Wikipedia entry.

3. Information Coverage in Multilingual Wikipedia

Wikipedia is a prominent example of collaborative knowledge generation and sharing. People contributing to Wikipedia articles' writing and verification are driven by a variety of reasons, including career, social, ideology, etc.[23] All Wikipedia articles are monitored by the Wikipedia community, which ensures the control of the quality and content of the articles.

Currently, Wikipedia has articles in more than 200 languages. The language that contains the largest number of articles is English,[12,24] but the size of non-English Wikipedia is growing fast and represents a rich corpus.[c]

Most current information selection applications that use Wikipedia as their corpus analyze Wikipedia articles written only in one language and assume that a Wikipedia article in any language is a reliable source of information. However, according to our observations, Wikipedia articles about the same entry (person, location, event, etc.) written in different languages frequently cover different sets of facts. Studying these differences can boost the development of various NLP applications (i.e. summarization, QA, new information detection, machine translation, etc.). According to our Wikipedia analysis, there are two major sources of differences in the articles about the same Wikipedia entry written in different languages:

- the amount of the information covered by the Wikipedia articles (the length of the Wikipedia articles);
- the choice of the information covered by the Wikipedia articles.

3.1. *Wikipedia article length difference*

The lengths of the Wikipedia articles about the same entry written in different languages vary. Perhaps, the amount of information in Wikipedia articles depends on how important a particular Wikipedia entry is for the community of people speaking in a particular language. In this work, the length of a Wikipedia article is measured in sentences used in the text description of a Wikipedia entry.

For example, baseball is popular in the USA, Latin America, and Japan but it is not widely spread in, for example, Europe or Africa. Thus, the articles about a legendary baseball player *Babe Ruth* exist in 26 languages.[d] The longest and the

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

[d]Wikipedia is changing constantly. The Wikipedia examples and data analyzed in this paper were collected on February 10, 2009, between 14:00 and 21:00 PST.

most detailed articles are the English and Japanese ones. The article about *Babe Ruth* in Croatian has one sentence, in Lithuanian — two sentences, in Finnish — six sentences. These short articles only list a few general biographical facts such as: date of birth, death; the fact that *Babe Ruth* was a legendary baseball player.

It is likely that the facts from the Lithuanian, Croatian, and Finnish articles about *Babe Ruth* (e.g. general biography facts) will be listed in a summary of the English language Wikipedia article about him. Moreover, all the sentences from the Lithuanian, Croatian, and Finnish articles about *Babe Ruth* have corresponding sentence in the English article about him and neither Lithuanian, nor Croatian, nor Finnish articles have information that is not covered in the English article.

3.2. *Choice of information*

At the same time, there exist Wikipedia entries whose shorter articles contain information facts that are not covered by longer articles. For example, the Wikipedia entry about *Isadora Duncan* has articles in 47 languages. The lengths of these articles vary greatly: from almost 150 sentences in the English language article to four sentences in Danish. *Isadora Duncan* was an American-born dancer who was very popular in Europe and was married to a Russian poet, *Sergey Esenin*. Certain facts (i.e. major biography dates) about *Isadora Duncan* are repeated in the articles in every language. However, shorter articles are not necessarily summaries of longer articles. For example, the article in Russian is almost four times shorter than the article in English, however, it contains information that is not covered in the article written in English. The same can be noted about the articles in French and Spanish.

4. Testing the Hypothesis about the Multilingual Wikipedia Information Overlap Structure: Summarization Experiment

Despite the differences in length and information coverage descried in Sec. 3, it is obvious that Wikipedia articles about the same entry written in different languages have a substantial amount of information that is common for articles in all (or in most) languages.

In our work, we suggest the hypothesis that the information overlap in multilingual Wikipedia corresponds well to the pyramid summarization evaluation model as well as the multidocument summarization information importance model. To prove this hypothesis we build a summarization system based on the presented multilingual Wikipedia information overlap hypothesis. We use this system to generate summaries of the English Wikipedia articles given the articles about the same entry in other languages.

4.1. *Summarization and wikipedia: Information overlap and information position*

Currently, information repetition (redundancy) is successfully used for both summary generation and summarization evaluation. The state-of-the-art (automatic

and manual) summarization evaluation approaches compare the automatically generated summaries against several model summaries. Such models are typically created manually. The more model summaries contain a certain information fact — the greater value it gets in the automatically generated summary.[1,2,25]

Many current successful multidocument summarization systems rely on the observation that the information facts that are good candidates to be included into the summary are likely to be repeated in many input documents.[3,4] At the same time, in the field of single document summarization there exists a consensus that taking the first several sentences is typically the best way to summarize a news article. This single news article summarization approach is often used as an evaluation baseline, called the *lead line* baseline, and is very difficult to beat.[26] The strength of this baseline is due to the rules that are typically applied by journalists to structure their news articles.

In contrast to news summarization, the lead section of Wikipedia articles can be used not as a baseline but rather as a human generated model summary. Wikipedia articles have a well-defined structure. Even though authors can write Wikipedia articles disregarding the established conventions, the crowdsourcing nature of Wikipedia leads toward the enforcement of the Wikipedia Manual of Style[e] for all Wikipedia articles. This phenomenon is successfully used by NLP systems to learn the typical structure of Wikipedia articles in general[27] and biographical Wikipedia articles in particular.[10] Specifically, the lead section[f] of Wikipedia articles "serves both as an introduction to the article and as a summary of its most important aspects. The lead should be able to stand alone as a concise overview of the article". In our experiment, we use this peculiarity of the Wikipedia article structure as the source of human generated model summaries.

In the summarization experiment that we set up to prove the correctness of our multilingual Wikipedia information overlap structure hypothesis we use the information repetition degree in the articles about the same Wikipedia entry written in different languages to identify those sentences from the English articles about this entry that are important. In addition, we rely on the typical structure of Wikipedia articles and output the identified important sentences in the order they appear in the English article. We use a sentence as a proxy for an information fact because most existing extractive multidocument summarizers use sentences as main generation units.

4.2. *Data set*

For our experiment, we use the list of people created for the Task 5 of DUC 2004: biography generation task (48 people).[g] First, we downloaded from Wikipedia all the articles in all the languages corresponding to each person from the DUC 2004

[e]http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style
[f]http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_(lead_section)
[g]http://duc.nist.gov/duc2004/tasks.html/

evaluation. We used Wikitext, the text that is used by Wikipedia authors and editors. Wikitext complies with the wiki markup language and can be processed by the Wikimedia content manager system into HTML which can then be viewed in a browser. This is the text that can be obtained through Wikipedia dumps.[h] For our experiment, 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.

For each Wikipedia entry (i.e. DUC 2004 person) we downloaded the corresponding articles in all the languages, including Esperanto, Latin, etc. We used the name of a person to find the article about this person in English and then we followed the links from the left side panel of the Wikipedia page template to get the articles in other languages. To facilitate the comparison of the articles written in different languages we used the Google machine translation tool[i] to translate the downloaded articles 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 to the language distribution in Wikipedia and the collected Wikipedia subset can be considered a representative sample.[5]

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

The Google machine translation system does not have publicly available evaluation numbers for all the language pairs. However, none of the freely available research systems cover as many languages as the Google machine translation system does. Given our information overlap hypothesis and the design of our summarization experiment we opt to use the Google machine translation system. It is more important for us to analyze Wikipedia articles in as many languages as possible, even though their translations into English might be poor, rather than have only a few high quality translations into English.

It must be noted that there exist summarization systems that specifically deal with the multilingual summarization task. Evans *et al.*[28] generate English summaries from English, Russian, and Japanese sources. Methods for identifying similarities and differences across English and Arabic news articles are studied in Ref. 29. Litvak *et al.*[30] work with two monolingual as well as with a bilingual English–Hebrew corpora. There exist many other systems that produce high quality

[h]http://download.wikimedia.org/
[i]http://translate.google.com/

summaries given input documents written in a specific set of languages. However, the main goal of our paper is to test the hypothesis that information overlap in the multilingual Wikipedia has a pyramid-like structure and thus, we need to use many more languages that are currently handled by the most successful multilingual multidocument summarizers. Also, the output summary of out system is a summary of a single document (Wikipedia article written in English) based on the information overlap among the articles about this entry currently existing in multilingual Wikipedia.

4.3. *Data processing tools*

After the Wikipedia articles about the DUC 2004 set were collected and translated, we divided these articles into sentences using the LingPipe sentence chunker.[31] For each DUC 2004 person we compared an article about this person in English against the articles about this person in the other languages that were handled by the Google translation system. We counted articles in how many languages had sentences corresponding to the sentences in the articles 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 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".[31] Before the similarity measure is applied to sentences, we eliminate all the stop words from the sentences. For our experiment, each sentence was treated as a separate document. The IDF value was computed based on the two articles under consideration (one — in English, the other one — translation into English). We used three similarity thresholds: 0.5, 0.35, 0.2.

4.4. *What was measured*

To evaluate how much information is repeated in the articles about the same person in different languages, we measure the similarity of the person's description in English and in other languages. Since each sentence is treated as a separate document, the number of tokens (words) for comparison is rather small. Thus, for the 0.5 similarity threshold, the sentences marked as similar are almost identical. The 0.35 and 0.2 thresholds allow us to search for nonidentical sentences that still have a substantial word overlap.

Our hypothesis is that those information facts (sentences) that are mentioned in the articles about a person in different languages fit well the pyramid summarization model. For example, if we are to summarize an article about a person from the English Wikipedia: first, we should add to the summary those sentences that have their counterparts in the most number of articles about this person in the languages other than English. Sentences added on this step correspond to the top level of the pyramid and represent the most important part of the English article about the

Table 1. Algorithm outline.

	Algorithm
1.	Submit the person's name to Wikipedia
2.	Get Wikipedia articles about this person in all languages
3.	Remove nonplain text information from the articles
4.	For all the languages handled by the Google MT, translate the articles from these languages into English
5.	Break English articles (the original English article and the translations into English from other languages) into sentences
6.	Identify what sentences from the original English-language article have similar sentences in the articles about the same person written in other languages
7.	Rank all the sentences from the original English-language article according to the number of languages that have similar sentences
8.	If several sentences are placed on the same level (have the same rank), list these sentences in the order they appear in the original English-language Wikipedia article
9.	Use the top three levels from the above ranking

person under analysis. If the length of the summary is not exhausted, then on the next step, we add to the summary the sentences 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 of 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.

4.5. *Output example and experiment discussion*

Table 2 presents three-level summaries for the English Wikipedia article about *Gene Autry*. Wikipedia has articles about *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.

When using a similarity of 0.5 we get one sentence from the English article that has counterparts in at least two other languages (here, in three other languages). This sentence's ID is 0: it is the first, introductory sentence of the English Wikipedia article about *Gene Autry*.

When using a 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 seven languages, rather than in three. The next level consists of the sentences from the English article 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.

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

#	Lang.	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

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 articles 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.

In the experiment described in this paper, we analyze only the sentences from the English text that appear in at least two other languages Thus, using the DUC 2004 set we created:

- for the similarity score of 0.2
  - — **one**-level summaries for two people[j];
  - — **two**-level summaries for two people[k];
  - — **three**-level summaries for 38 people.[l]
- for the similarity score of 0.35
  - — **one**-level summaries for three people;
  - — **two**-level summaries for four people;
  - — **three**-level summaries for 35 people.
- for the similarity score of 0.5
  - — **one**-level summaries for five people;
  - — **two**-level summaries for eight people;
  - — **three**-level summaries for 25 people.

From the above numbers it is clear that not all the people under analysis have articles in two languages other than English with sentences that are similar to the sentences in the English article. Given that on the sentence level the similarity of 0.5 corresponds to almost identical sentences, we fail to generate summaries for several people.

The above results should be read as follows: given the similarity of 0.2, we can create summaries for 42 people: 38 summaries have sentences on all three levels, two summaries have sentences only on the first two levels, and the remaining two summaries have sentences only on the first level. The length of the created summaries is measured in sentences. Table 3 presents information about the average number of sentences on each level.

Table 3. Average number of sentences for each level.

Similarity Measure	Level One	Level Two	Level Three
0.2	1	2.2	2.42
0.35	1	1.6	2
0.5	1	7	6

[j]**One**- level summary: only one level (Level 1) is filled in.
[k]**Two**- level summary: two levels (Level 1 and 2) are filled in.
[l]**Three**-level summary: all three levels (Level 1, 2 and 3) are filled in.

4.6. *Model summaries*

To prove the correctness of our hypothesis, in addition to the summaries generated following the algorithm outlined in Table 1, we create model lead line summaries for the English Wikipedia articles about the DUC 2004 people. We use three sets of model summaries created for each of the similarity measures (0.2, 0.35, 0.5). For a given person, similarity measure, and summary level, we extract from the English Wikipedia article about the person under analysis as many sentences as are added to the respective summary level by our summarization system based on the multilingual Wikipedia information overlap structure hypothesis.

As described in Sec. 4.2, to analyze the information overlap in Multilingual Wikipedia we use pre-processed Wikitext. To generate model summaries we do not use this pre-processed text, rather we manually extract the first sentences from the Wikipedia articles about the DUC 2004 people written in English. As described in Sec. 4.1, the lead section of a Wikipedia article can be used as the summary of this article. However, according to the above procedure, the model summaries can contain only a part of the lead section.

Table 4 presents the model (lead section) three-level summaries for the English Wikipedia article about *Gene Autry*. The number of sentences on each level is the same as the number of sentences on the respective levels of summaries presented in Table 2.

4.7. *Evaluation procedure*

To prove the validity of our hypothesis, we evaluate the quality of the summaries produced by our system that exploits information overall in multilingual Wikipedia and compare it to the quality of the model summaries. We used Amazon Mechanical Turk[m] as the source of human subjects who can reliably evaluate certain NLP tasks.[32] All in all we wanted to evaluate 258 outputs: six summaries for each of the 43 people from the DUC 2004 set, three automatically generated summaries (one — for each similarity measure), and three corresponding model summaries. For each of the 258 outputs we recruited five human annotators. The annotators were provided with: the name of the person; link to the English-language Wikipedia article about this person; three-level summary of this Wikipedia article. The annotators did not know whether they got an automatically generated or a model summary. 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 article 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 article 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?

[m]http://www.mturk.com

Table 4. Three-level summaries for Gene Autry consisting of the lead sentences. The number of sentences on each level is the same as the number of sentences on the respective levels of summaries presented in Table 2.

#	ID	Text
Similarity 0.5		
1	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	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	1	Autry, the grandson of a Methodist preacher, was born near Tioga, Texas
	2	His parents, Delbert Autry and Elnora Ozment, moved to Oklahoma in the 1920s
3	3	After leaving high school in 1925, Autry worked as a telegrapher for the St. Louis-San Francisco Railway
	4	Talent with the guitar and his voice led to performing at local dances
	5	After an encouraging chance encounter with Will Rogers, he began performing on local radio in 1928 as "Oklahoma's Yodeling Cowboy"
Similarity 0.2		
1	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	1	Autry, the grandson of a Methodist preacher, was born near Tioga, Texas
	2	His parents, Delbert Autry and Elnora Ozment, moved to Oklahoma in the 1920s
3	3	After leaving high school in 1925, Autry worked as a telegrapher for the St. Louis-San Francisco Railway
	4	Talent with the guitar and his voice led to performing at local dances

- Assume that the summary of the Wikipedia article 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. Table 5 shows the percentage of summaries that are considered *good* on each level by the majority of the annotators (at least three out of five) for both model summaries and summaries that are created relying on the hypothesis about the pyramid-structure of the information overlap in multilingual Wikipedia. It is possible to see that the scores of the system generated summaries based on our hypothesis are lower than the scores for the human generated models. However, we would like to emphasize that:

- Lead section for Wikipedia articles are human generated summaries that are constantly checked and approved by Wikipedia editors;
- Lead section model summaries used in our experiment are created by manually extracting the first sentences from Wikipedia articles, while system-generated

Table 5. The percentage of the DUC 2004 people whose summaries are considered good by the majority of MTurk annotators.

Similarity Measure	System or Model	Three-Level Summary	Two-Level Summary	One-Level Summary
0.2	model	97.62%	100%	97.37%
	system	76.19%	87.5%	84.21%
0.35	model	92.90%	87.18%	80.00%
	system	81.00%	71.80%	71.43%
0.5	model	94.74%	90.91%	96.00%
	system	60.53%	57.58%	80.00%

summaries consist of the sentences that are extracted from the automatically processed text from the *Edit* Wikipedia tab.

We believe that despite the fact that the summaries produced by our summarization system get lower scores than the model summaries, the numbers presented in Table 5 can be used as a proof of the validity of our multilingual Wikipedia information overlap structure hypothesis. In all the cases the summaries built according to this hypothesis are considered reasonably good by the majority of the human annotators. In Sec. 4.8, we present the error analysis that shows that many of the automatically-generated summaries' weak points can be avoided by using more sophisticated text processing tools.

4.8. *Error analysis*

As described in Sec. 4.2, we use Wikitext to analyze the information overlap in Multilingual Wikipedia. Wikitext is the text that is used by Wikipedia authors and editors. There is no commonly accepted standard Wikitext language, thus, our final text had a certain amount of noise. According to our analysis, substantial number of summaries that were judged as being *bad* contained sentences that were affected by the pre-processing of the Wikitext data.

After analyzing the sentences included into the system-generated summaries, we detected four types of problems that could cause the negative rating of the produced summaries: broken sentences; weak content; poor context; redundant information. Table 6 summarizes our error analysis and contains the frequency of each problem type.

4.8.1. *Error type* 1: *Broken sentences*

According to our analysis, several summaries contain truncated sentences due to incorrect sentence chunking. Let us analyze the following sentence from the English-language Wikipedia article about *Paul Wellstone*:

> [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

Table 6. Error analysis.

Level # (# of sentences)	Broken Sentence	Weak Content	Poor Context	Redundant Sentence
Similarity 0.5				
Level 1 (38)	3	6	2	0
Level 2 (56)	17	14	3	0
Level 3 (50)	7	21	8	1
Similarity 0.35				
Level 1 (40)	3	4	0	0
Level 2 (57)	7	13	1	1
Level 3 (55)	5	15	4	5
Similarity 0.2				
Level 1 (42)	4	3	1	0
Level 2 (87)	17	9	10	9
Level 3 (92)	25	13	5	1

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 output summary. Despite the fact that this portion contains important biographical information, it cannot be used as a stand-alone sentence. This portion of the sentence was used as the only sentences on Level 1 and thus, obviously, the summary consisting only of one level was judged as *bad* by the majority of the annotators.

We call such sentences *broken* as they do not flow comprehensibly because a period used for an abbreviation is treated as a full stop. This results in sentences ending at a person's middle initial, title, or any acronym involving periods. Consequently, a piece of text that we erroneously treat as a sentence is only a part of a sentence: typically, either the beginning or the end of a sentence. The pieces of text either finish abruptly or begin awkwardly. Our corpus has a few cases where such an awkward continuity of a single sentence goes on to take up to three or four text pieces that are added to the output as separate sentences, thus increasing the number of broken sentences in the output.

4.8.2. *Error type* 2: *Weak content*

Determining the quality of a sentence's content is subjective, since different users could be searching for different details.[1,2] When a number of sentences included into the summary is dramatically smaller than the article itself, this issue becomes even more profound. For example, an article can contain a wide variety of details on many topics, however, each sentence contains information only on one of the topics. Thus, if the number of sentences is considerably smaller than the number of topics covered in the article, it is obvious that many topics will not be covered in the summary.

For example, one of the MTurk annotators who rated the summary about *Gene Autry* (Table 2: similarity measure of 0.35) notes that

> "The summary includes too many song references and fails to mention Autry's hall of fame induction".

Another example of what can be considered as weak content can be found in the summary about *Hugo Chávez* constructed using the similarity measure of 0.35:

- Level 1
  (1) Early life (1954–1992) Chávez was born on July 28, 1954 in the town of Sabaneta, Barinas.
- Level 2
  (1) After a two-year imprisonment, Chávez was pardoned by President Rafael Caldera in 1994.
- Level 3
  (1) Chávez went on to win the 1998 presidential election on December 6, 1998 with 56% of the votes.

The annotators' comments submitted for this summary are listed below. They show that though the summary contains useful information, this information is not enough for a summary to be considered good. Perhaps, in future work this obstacle can be overcome by allowing more sentences to be included in the output summary.

- "There is no context between the three sentences. It does not say what he is famous or notable for. He was imprisoned... for what? He won the presidency... of which country?"
- "Very thin summary".

Here is how the MTurk annotators rated the above summary about *Hugo Chávez*:

- Level 1: Good - 1; Bad - 4;
- Level 2: Good - 0; Bad - 5;
- Level 3: Good - 1; Bad - 4.

4.8.3. *Error type* 3: *Poor context*

The problem of poor context is related to the problem of weak content. In several cases the problem of poor context can be eliminated by simple re-ordering of sentences.

Below is the summary output produced by our system with the similarity of 0.2 for the articles about *Thabo Mbeki*. Only three sentences are included into the summary (one for each level).

- Level 1
  (1) In November 2008, "The New York Times" reported that due to Thabo Mbeki's rejection of scientific consensus on AIDS and his embrace of AIDS denialism, an estimated 365,000 people perished in South Africa.
- Level 2
  (1) His father was a stalwart of the African National Congress (ANC) and the South African Communist Party.
- Level 3
  (1) Thabo Mbeki was the executive face of government in South Africa from 1994.

Three of the five annotators left comments concerning the quality of the output for the article about *Thabo Mbeki*:

- "The summaries don't work well in their starts. Start with presidential, then notable aids epidemic then father figure".
- "The sentence listed in Level 3 is the only one which provides detail on who Thabo Mbeki is. The Level 1 and Level 2 summary describes the subject's actions, belief and relations".
- "I think the information would better be presented in reverse, with the sentence from Level 3 at the start, followed by the sentence from Level 2, and concluding with the sentence from Level 1".

Here is how MTurk annotators rated the above summary about *Thabo Mbeki*:

- Level 1: Good - 1; Bad - 4;
- Level 2: Good - 2; Bad - 3;
- Level 3: Good - 3; Bad - 2.

Obviously, the sentences included into the summary contain information worth summarizing; however, their placement relative to the rest of the output sentences rendered it incomprehensible. This usually occurs when a sentence is chosen from the center of an article's content, rather than being the first sentence of an article's new section.

According to the annotators' comments, a simple re-ordering of the levels could have improved the rating given to the summary. We believe that in our future experiments this problem could be fixed by using the technique developed for identifying the most likely order of information in a Wikipedia article that is described in Ref. 27.

4.8.4. *Error type* 4: *Redundant information*

In our system we do not have any control for the redundancy of the information. According to Table 6, information overlap in the sentences within the same article does not happen often. In our future experiments we plan on using of the algorithms developed specifically to avoid redundancy in summarization systems.[4,33]

4.8.5. *Error analysis: Conclusion*

Overall, the quality of the summaries generated according to our information overlap hypothesis is lower than the quality of the model summaries. However, many of the automatically generated summaries' errors can be eliminated by using more sophisticated text processing tools. Thus, we believe that the quality of the automatically generated summaries is sufficient to proof the correctness of our hypothesis.

5. Conclusion and Future Work

In this paper we study the structure of information overlap in multilingual Wikipedia. We introduce and prove the hypothesis that the structure of this information overlap is similar to the information overlap structure (pyramid model) used in summarization evaluation, as well as the information overlap/repetition structure used to identify important information for multidocument summarization.

We believe that the understanding of the information overlap structure in multilingual Wikipedia can be used in a variety of ways. For example, it can be used for automatic generation of short, place-holder entry descriptions in new languages. Second, the pyramid structure of information overlap can be used for judging the trustworthiness of information facts mentioned in Wikipedia articles: the more important (and the more trusted) the information fact for a particular Wikipedia entry is, the higher the chances that this information is covered (repeated) in the articles in many languages.

While in this paper we focus on studying the structure of information overlap, in our future work, we are also interested in investigating information asymmetries in multilingual Wikipedia. Highlighting differences in Wikipedia articles about the same entry written in different languages can be used for suggesting automatic updates of Wikipedia articles, information tracking, information trustworthiness evaluation, etc. Information asymmetries is an interesting topic to study given that the choice of information can reveal the authors attitude toward the Wikipedia entry, especially in controversial topics.

References

1. S. Teufel and H. V. Halteren, Evaluating information content by factoid analysis: Human annotation and stability, in *Proc. 42th Annual Meeting of the Association for Computational Linguistics* (*ACL 2004*), Barcelona, Spain, 2004, pp. 419–426.
2. A. Nenkova, R. Passonneau and K. McKeown, The pyramid method: Incorporating human content selection variation in summarization evaluation, *ACM Trans. Speech Lang. Process.* **4**(2) (2007) 4.
3. E. Filatova and V. Hatzivassiloglou, A formal model for information selection in multi-sentence text extraction, in *Proc. 20th Int. Conf. Computational Linguistics* (*COLING 2004*), Geneva, Switzerland, 2004.
4. D. Radev, H. Jing, M. Styś and D. Tam, Centroid-based summarization of multiple documents, *Inf. Process. Manage.* **40**(6) (2004) 919–938.

5. E. Filatova, Directions for exploiting asymmetries in multilingual Wikipedia, in *Proc. NAACL Workshop on Cross Lingual Information Access*, Boulder, CO, USA, 2009.
6. E. Filatova, Multilingual Wikipedia, summarization, and information trustworthiness, in *Proc. NAACL Workshop on Information Access in a Multilingual World*, Boston, MA, USA, 2009.
7. D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke and S. Schlobach, Using Wikipedia at the TREC QA track, in *Proc. Text REtrieval Conf.* (*TREC 2004*), 2004.
8. D. Buscaldi and P. Rosso, Mining knowledge from Wikipedia for the question answering task, in *Proc. Fifth Int. Conf. Language Resources and Evaluation* (*LREC 2006*), Genoa, Italy, 2006.
9. J. Ko, T. Mitamura and E. Nyberg, Language-independent probabilistic answer ranking for multilingual question answering, in *Proc. 45th Annual Meeting of the Association for Computational Linguistics* (*ACL 2007*), Prague, Czech Republic, 2007.
10. F. Baidsy, J. Hirschberg and E. Filatova, An unsupervised approach to biography production using Wikipedia, in *Proc. 46th Annual Meeting of the Association for Computational Linguistics* (*ACL 2008*), Columbus, OH, USA, 2008.
11. R. Nelken and E. Yamangil, Mining Wikipedia's article revision history for training computational linguistics algorithms, in *Proc. AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, Chicago, IL, USA, 2008.
12. S. F. Adafre and M. de Rijke, Finding similar sentences across multiple languages in Wikipedia, in *Proc. Conf. European Chapter of the Association for Computational Linguistics, Workshop on New Text — Wikis and Blogs and other Dynamic Text Sources*, Trento, Italy, 2006.
13. A. Richman and P. Schone, Mining Wiki resources for multilingual named entity recognition, in *Proc. 46th Annual Meeting of the Association for Computational Linguistics* (*ACL 2008*), Columbus, OH, USA, 2008.
14. P. Schönhofen, A. Benczúr, I. Bíró and K. Csalogány, Performing cross-language retrieval with Wikipedia, in *Proc. Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 2007.
15. S. Ferrández, A. Toral, Ó. Ferrández, A. Ferrández and R. Munoz, Applying Wikipedia's multilingual knowledge to cross-lingual Question Answering, *Lecture Notes in Computer Science* (*LNCS*)*: Natural Language Processing and Information Systems*, Vol. 4592, (2007), pp. 352–363.
16. B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye and V. Raman, Assigning trust to Wikipedia content, in *Proc. Int. Symp. Wikis* (*WikiSym 2008*), Porto, Portugal, 2008.
17. H. Zeng, M. Alhossaini, L. Ding, R. Fikes and D. McGuinness, Computing trust from revision history, in *Proc. Int. Conf. Privacy, Security and Trust* (*PST 2006*), Markham, Canada, 2006.
18. D. McGuinness, H. Zeng, P. P. da Silva, L. Ding, D. Narayanan and M. Bhaowal, Investigations into trust for collaborative information repositories: A Wikipedia case study, in *Proc. Workshop on the Models of Trust for the Web* (*MTW 2006*), Edinburgh, UK, 2006.
19. K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay and E. Eskin, Towards multidocument summarization by reformulation: Progress and prospects, in *Proc. 16th National Conf. American Association for Artificial Intelligence* (*AAAI 1999*), Orlando, Florida, 1999, pp. 453–460.
20. R. Barzilay, N. Elhadad and K. McKeown, Inferring strategies for sentence ordering in multidocument news summarization, *J. Artif. Intell. Res.* **17** (2002) 35–55.

21. S. Harabagiu, F. Lăcătuşu and S. Maiorano, Multi-document summaries based on semantic redundancy, in *Proc. 14th Florida AI Conf.* (*FLAIRS 2003*), St. Augustine, Florida, 2003.
22. E. Adar, M. Skinner and D. Weld, Information arbitrage in multi-lingual Wikipedia, in *Proc. Second ACM Int. Conf. Web Search and Data Mining*, Barcelona, Spain, 2009.
23. O. Nov, What motivates Wikipedians? *Commun. ACM* **50**(11) (2007) 60–64.
24. L. Denoyer and P. Gallinari, The Wikipedia XML Corpus, SIGIR Forum.
25. C.-Y. Lin and E. Hovy, Automatic evaluation of summaries using N-gram co-occurrence statistics, in *Proc. Language Technology Conf.* (*HLT-NAACL 2003*), Edmonton, Canada, 2003.
26. P. Over and W. Liggett, Introduction to DUC-2002: An intrinsic evaluation of generic news text summarization systems, in *Proc. Workshop on Automatic Summarization* (*DUC 2002*), Philadelphia, PA, USA, 2002.
27. C. Sauper and R. Barzilay, Automatically generating Wikipedia articles: A structure-aware approach, in *Proc. 47th Annual Meeting of the Association for Computational Linguistics* (*ACL 2009*), Singapore, 2009.
28. D. Evans, J. Klavans and K. McKeown, Columbia Newsblaster: Multilingual news summarization on the web, in *Demonstration Papers at HLT-NAACL 2004, HLT-NAACL–Demonstrations '04, Association for Computational Linguistics*, Stroudsburg, PA, USA, 2004, pp. 1–4.
29. D. Evans and K. McKeown, Identifying similarities and differences across English and Arabic news, in *Int. Conf. Intelligence Analysis*, McLean, VA, USA, 2005.
30. M. Litvak, M. Last and M. Friedman, A new approach to improving multilingual summarization using a genetic algorithm, in *Proc. 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 927–936.
31. Alias-i, Lingpipe 3.7.0. (accessed January 19, 2009), http://alias-i.com/lingpipe (2009).
32. V. Sheng, F. Provost and P. Ipeirotis, Get another label? Improving data quality and data mining using multiple, noisy labelers, in *Proc. Fourteenth ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (*KDD 2008*), Las Vegas, NV, USA, 2008.
33. J. Carbonell and J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, in *Proc. 21nd Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval* (*SIGIR 1998*), New York, NY, USA, 1998, pp. 335–336.