Finding Similar Sentences across Multiple Languages in Wikipedia

Sisay Fissaha Adafre **Maarten de Rijke**
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
sfissaha,mdr@science.uva.nl

Abstract

We investigate whether the Wikipedia corpus is amenable to multilingual analysis that aims at generating parallel corpora. We present the results of the application of two simple heuristics for the identification of similar text across multiple languages in Wikipedia. Despite the simplicity of the methods, evaluation carried out on a sample of Wikipedia pages shows encouraging results.

1 Introduction

Parallel corpora form the basis of much multilingual research in natural language processing, ranging from developing multilingual lexicons to statistical machine translation systems. As a consequence, collecting and aligning text corpora written in different languages constitutes an important prerequisite for these research activities.

Wikipedia is a multilingual free online encyclopedia. Currently, it has entries for more than 200 languages, the English Wikipedia being the largest one with 895,674 articles, and no fewer than eight language versions having upwards of 100,000 articles as of January 2006. As can be seen in Figure 1, Wikipedia pages for major European languages have reached a level where they can support multilingual research. Despite these developments in its content, research on Wikipedia has largely focused on monolingual aspects so far; see e.g., (Voss, 2005) for an overview.

In this paper, we focus on multilingual aspects of Wikipedia. Particularly, we investigate to what extent we can use properties of Wikipedia itself to generate similar sentences acrose different languages. As usual, we consider two sentences similar if they contain (some or a large amount of) overlapping information. This includes cases in which sentences may be exact translations of each other, one sentence may be contained within another, or both share some bits of information.

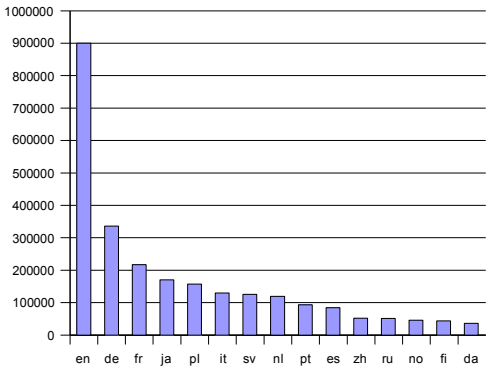

Figure 1: Wikipedia pages for the top 15 languages

The conceptually simple but fundamental task of identifying similar sentences across multiple languages has a number of motivations. For a start, and as mentioned earlier, sentence aligned corpora play an important role in corpus based language processing methods in general. Second, in the context of Wikipedia, being able to align similar sentences across multiple languages provides insight into Wikipedia as a knowledge source: to which extent does a given topic get different kinds of attention in different languages? And thirdly, the ability to find similar content in other languages while creating a page for a topic in one language constitutes a useful type of editing support. Furthermore, finding similar content acrose different languages can form the basis for multilingual summarization and question answering support for

Wikipedia; at present the latter task is being developed into a pilot for CLEF 2006 (WiQA, 2006).

There are different approaches for finding similar sentences across multiple languages in non-parallel but comparable corpora. Most methods for finding similar sentences assume the availability of a clean parallel corpus. In Wikipedia, two versions of a Wikipedia topic in two different languages are a good starting point for searching similar sentences. However, these pages may not always conform to the typical definitions of a bitext which current techniques assume. Bitext generally refers to two versions of a text in two different languages (Melamed, 1996). Though it is not known how information is shared among the different languages in Wikipedia, some pages tend to be translations of each other whereas the majority of the pages tend to be written independently of each other. Therefore, two versions of the same topic in two different languages can not simply be taken as parallel corpora. This in turn limits the application of some of the currently available techniques.

In this paper, we present two approaches for finding similar sentences across multiple languages in Wikipedia. The first approach uses freely available online machine translation resources for translating pages and then carries out monolingual sentence similarity. The approach needs a translation system, and these are not available for every pair of languages in Wikipedia.

This motivates a second approach to finding similar sentences across multiple languages, one which uses a bilingual title translation lexicon induced automatically using the link structure of Wikipedia. Briefly, two sentences are similar if they link to the same entities (or rather: to pages about the same entities), and we use Wikipedia itself to relate pages about a given entity across multiple languages. In Wikipedia, pages on the same topic in different languages are topically closely related. This means that even if one page is not a translation of another, they tend to share some common information. Our underlying assumption here is that there is a general agreement on the kind of information that needs to be included in the pages of different types of topics such as a biography of a person, and the definition and description of a concept etc., and that this agreement is to a consderable extent "materialized" in the hypertext links (and their anchor texts) in Wikipedia.

Our main research question in this paper is this: how do the two methods just outlined differ? A priori it seems that the translation based approach to finding similar sentences across multiple languages will have a higher recall than the link-based method, while the latter outperforms the former in terms of precision. Is this correct?

The remainder of the paper is organized as follows. In Section 2, we briefly discuss related work. Section 3 provides a detailed description of Wikipedia as a corpus. The two approaches to identifying similar sentences across multiple languages are presented in Section 4. An experimental evaluation is presented in Section 5. We conclude in Section 6.

2 Related Work

The main focus of this paper lies with multilingual text similarity and its application to information access in the context of Wikipedia. Current research work related to Wikipedia mostly describes its monolingual properties (Ciffolilli, 2003; Viégas et al., 2004; Lih, 2004; Miller, 2005; Bellomi and Bonato, 2005; Voss, 2005; Fissaha Adafre and de Rijke, 2005). This is probably due to the fact that different language versions of Wikipedia have different growth rates. Others describe its application in question answering and other types of IR systems (Ahn et al., 2005). We believe that currently, Wikipedia pages for major European languages have reached a level where they can support multilingual research.

On the other hand, there is a rich body of knowledge relating to multilingual text similarity. These include example-based machine translation, cross-lingual information retrieval, statistical machine translation, sentence alignment cost functions, and bilingual phrase translation (Kirk Evans, 2005). Each approach uses relatively different features (content and structural features) in identifying similar text from bilingual corpora. Furthermore, most methods assume that the bilingual corpora can be sentence aligned. This assumption does not hold for our case since our corpus is not parallel. In this paper, we use content based features for identifying similar text across multilingual corpora. Particularly, we compare bilingual lexicon and MT system based methods for identifying similar text in Wikipedia.

3 Wikipedia as a Multilingual Corpus

Wikipedia is a free online encyclopedia which is administered by the non-profit Wikimedia Foundation. The aim of the project is to develop free encyclopedias for different languages. It is a collaborative effort of a community of volunteers, and its content can be edited by anyone. It is attracting increasing attention amongst web users and has joined the top 50 most popular sites.

As of January 1, 2006, there are versions of Wikipedia in more than 200 languages, with sizes ranging from 1 to over 800,000 articles. We used the ascii text version of the English and Dutch Wikipedia, which are available as database dumps. Each entry of the encyclopedia (a page in the online version) corresponds to a single line in the text file. Each line consists of an ID (usually the name of the entity) followed by its description. The description part contains the body of the text that describes the entity. It contains a mixture of plain text and text with html tags. References to other Wikipedia pages in the text are marked using "[[" "]]" which corresponds to a hyperlink on the online version of Wikipedia. Most of the formatting information which is not relevant for the current task has been removed.

3.1 Links within a single language

Wikipedia is a hypertext document with a rich link structure. A description of an entity usually contains hypertext links to other pages within or outside Wikipedia. The majority of these links correspond to entities, which are related to the entity being described, and have a separate entry in Wikipedia. These links are used to guide the reader to a more detailed description of the concept denoted by the anchor text. In other words, the links in Wikipedia typically indicate a topical association between the pages, or rather the entities being described by the pages. E.g., in describing a particular person, reference will be made to such entities as country, organization and other important entities which are related to it and which themselves have entries in Wikipedia. In general, due to the peculiar characteristics of an encyclopedia corpus, the hyperlinks found in encyclopedia text are used to exemplify those instances of hyperlinks that exist among topically related entities (Ghani et al., 2001; Rao and Turoff, 1990).

Each Wikipedia page is identified with a unique ID. These IDs are formed by concatenating the words of the titles of the Wikipedia pages which are unique for each page, e.g., the page on Vincent van Gogh has "Vincent van Gogh" as its title and "Vincent_van_Gogh" as its ID. Each page may, however, be represented by different anchor texts in a hyperlink. The anchor texts may be simple morphological variants of the title such as plural form or may represent closely related semantic concept. For example, the anchor text "Dutch" may point to the page for the Netherlands. In a sense, the IDs function as the canonical form for several related concepts.

3.2 Links across different languages

Different versions of a page in different languages are also hyperlinked. For a given page, translations of its title in other languages for which pages exist are given as hyperlinks. This property is particularly useful for the current task as it helps us to align the corpus at the page level. Furthermore, it also allows us to induce bilingual lexicon consisting of the Wikipedia titles. Conceptual mismatch between the pages (e.g. *Roof* vs *Dakconstructie*) is rare, and the lexicon is generally of high quality. Unlike the general lexicon, this lexicon contains a relatively large number of names of individuals and other entities which are highly informative and hence are useful in identifying similar text. This lexicon will form the backbone of one of the methods for identifying similar text across different languages, as will be shown in Section 4.

4 Approaches

We describe two approaches for identifying similar sentences across different languages. The first uses an MT system to obtain a rough translation of a given page in one language into another and then uses word overlap between sentences as a similarity measure. One advantage of this method is that it relies on a large lexical resource which is bigger than what can be extracted from Wikipedia. However, the translation can be less accurate especially for the Wikipedia titles which form part of the content of a page and are very informative.

The second approach relies on a bilingual lexicon which is generated from Wikipedia using the link structure: pages on the same topic in different languages are hyperlinked; see Figure 2. We use the titles of the pages that are linked in this manner to create a bilingual lexicon. Thus, our bilingual lexicon consists of terms that represent

concepts or entities that have entries in Wikipedia, and we will represent sentences by entries from this lexicon: an entry is used to represent the content of a sentence if the sentence contains a hypertext link to the Wikipedia page for that entry. Sentence similarity is then captured in terms of the shared lexicon entries they share. In other words, the similarity measure that we use in this approach is based on "concept" or "page title" overlap. Intuitively, this approach has the advantage of producing a brief but highly accurate representation of sentences, more accurate, we assume than the MT approach as the titles carry important semantic information; it will also be more accurate than the MT approach because the translations of the titles are done manually.

Figure 2: Links to pages devoted to the same topic in other languages.

Both approaches assume that the Wikipedia corpus is aligned at the page level. This is easily achieved using the link structure since, again, pages on the same topic in different languages are hyperlinked. This, in turns, narrows down the search for similar text to a page level. Hence, for a given text of a page (sentence or chunk) in one language, we search for its equivalent text (sentence or chunk) only in the corresponding page in the other language, not in the entire corpus.

We now describe the two approaches in more detail. To remain focused and avoid getting lost in technical details, we consider only two languages in our technical descriptions and evaluations below: Dutch and English; it will be clear from our presentation, however, that our second approach can be used for any pair of languages in Wikipedia.

4.1 An MT based approach

In this approach, we translate the Dutch Wikipedia page into English using an online MT system. We refer to the English page as *source* and the translated (Dutch page) version as *target*. We used the Babelfish MT system of Altavista. It supports a number of language pairs among which are Dutch-English pairs. After both pages have been made available in English, we split the pages into sentences or text chucks. We then link each text chunk or sentence in the *source* to each chuck or sentence in the *target*. Following this we compute a simple word overlap score for each pair. We used the Jaccard similarity measure for this purpose. Content words are our main features for the computation of similarity, hence, we remove stopwords. Grammatically correct translations may not be necessary since we are using simple word overlap as our similarity measure.

The above procedure will generate a large set of pairs, not all of which will actually be similar. Therefore, we filter the list assuming a one-to-one correspondence, where for each source sentence we identify at most one target sentence. This is a rather strict criterion (another possibility being one-to-many), given the fact that the corpus is generally assumed to be not parallel. But it gives some idea on how much of the text corpus can be aligned at smaller units (i.e., sentence or text chunks).

Filtering works as follows. First we sort the pairs in decreasing order of their similarity scores. This results in a ranked list of text pairs in which the most similar pairs are ranked top whereas the least similar pairs are ranked bottom. Next we take the top most ranking pair. Since we are assuming a one-to-one correspondence, we remove all other pairs ranked lower in the list containing either of the the sentences or text chunks in the top ranking pair. We then repeat this process taking the second top ranking pair. Each step results in a smaller list. The process continues until there is no more pair to remove.

4.2 Using a link-based bilingual lexicon

As mentioned previously, this approach makes use of a bilingual lexicon that is generated from Wikipedia using the link structure. A high level description of the algorithm is given in Figure 3. Below, we first describe how the bilingual lexicon is acquired and how it is used for enriching the link structure of Wikipedia. Finally, we detail how the

- Generating bilingual lexicon
- Given a topic, get the corresponding pages from English and Dutch Wikipedia
- Split pages into sentences and enrich the hyperlinks in the sentence or identify named-entities in the pages.
- Represent the sentences in these pages using the bilingual lexicon.
- Compute term overlap between the sentences thus represented.

Figure 3: The Pseudo-algorithm for identifying similar sentences using a link-based bilingual lexicon.

bilingual lexicon is used for the identification of similar sentences.

Generating the bilingual lexicon

Unlike the MT based approach, which uses content words from the general vocabulary as features, in this approach, we use page titles and their translations (as obtained through hyperlinks as explained above) as our primitives for the computation of multilingual similarity. The first step of this approach, then, is acquiring the bilingual lexicon, but this is relatively straightforward. For each Wikipedia page in one language, translations of the title in other languages, for which there are separate entries, are given as hyperlinks. This information is used to generate a bilingual translation lexicon. Most of these titles are content bearing noun phrases and are very useful in multilingual similarity computation (Kirk Evans, 2005). Most of these noun phrases are already disambuiguated, and may consist of either a single word or multiword units.

Wikipedia uses a redirection facility to map several titles into a canonical form. These titles are mostly synonymous expressions. We used Wikipedia's redirect feature to identify synonymous expression.

Canonical representation of a sentence

Once we have the bilingual lexicon, the next step is to represent the sentences in both language pairs using this lexicon. Each sentence is represented by the set of hyperlinks it contains. We search each hyperlink in the bilingual lexicon. If it is found, we replace the hyperlink with the corresponding unique identification of the bilingual lexicon entry. If it is not found, the hyperlink will be included as is as part of the representation. This is done since Dutch and English are closely related languages and may share many cognate pairs.

Enriching the Wikipedia link structure

As described in the previous section, the method uses hyperlinks in a sentence as a highly focused entity-based representation of the aboutness of the sentence. In Wikipedia, not all occurrences of named-entities or concepts that have entries in Wikipedia are actually used as anchor text of a hypertext link; because of this, a number of sentences may needlessly be left out from the similarity computation process. In order to avoid this problem, we automatically identify other relevant hyperlinks using the bilingual lexicon generated in the previous section.

Identification of additional hyperlinks in Wikipedia sentences works as follows. First we split the sentences into constituent words. We then generate N gram words keeping the relative order of words in the sentences. Since the anchor texts of hypertext links may be multiword expressions, we start with higher order N gram words (N=4). We search these N grams in the bilingual lexicon. If the N gram is found in the lexicon, it is taken as a new hyperlink and will form part of the representation of a sentence. The process is repeated for lower order N grams.

Identifying similar sentences

Once we are done representing the sentences as described previously, the final step involves computation of the term overlap between the sentence pairs and filtering the resulting list. The remaining steps are similar to those described in the MT based approach. For completeness, we briefly repeat the steps here. First, all sentences from a Dutch Wikipedia page are linked to all sentences of the corresponding English Wikipedia page. We then compute the similarity between the sentence representations, using the Jaccard similarity coefficient.

A sentence in Dutch page may be similar to several sentences in English page which may result in a large number of spurious pairs. Therefore, we filter the list using the following recursive procedure. First, the sentence pairs are sorted by their similarity scores. We take the pairs with the highest similarity scores. We then eliminate all

other sentence pairs from the list that contain either of sentences in this pair. We continue this process taking the second highest ranking pair. Note that this procedure assumes a one-to-one matching rule; a sentences in Dutch can be linked to at most one sentence in English.

5 Experimental Evaluation

Now that we have described the two algorithms for identifying similar sentences, we return to our research questions. In order to answer them we run the experiment described below.

5.1 Set-up

We took a random sample of 30 English-Dutch Wikipedia page pairs. Each page is split into sentences. We generated candidate Dutch-English sentence pairs and passed them on to the two methods. Both methods return a ranked list of sentence pairs that are similar. As explained above, we assumed a one-to-one correspondence, i.e., one English sentence can be linked to at most to one Dutch sentence.

The outputs of the systems are manually evaluated. We apply a relatively lenient criteria in assessing the results. If two sentences overlap interms of their information content then we consider them to be similar. This includes cases in which sentences may be exact translation of each other, one sentence may be contained within another, or both share some bits of information.

5.2 Results

Table 1 shows the results of the two methods described in Section 4. In the table, we give two types of numbers for each of the two methods *MT* and *Bilingual lexicon*: *Total* (the total number of sentence pairs) and *Match* (the number of correctly identified sentence pairs) generated by the two approaches.

Overall, the two approaches tend to produce similar numbers of correctly identified similar sentence pairs. The systems seem to perform well on pages which tend to be alignable at sentence level, i.e., parallel. This is clearly seen on the following pages: *Pierluigi Collina*, *Marcus Cornelius Fronto*, *George F. Kennan*, which show a high similarity at sentence level. Some pages contain very small description and hence the figures for correct similar sentences are also small. Other topics such as *Classicism* (Dutch: *Classicisme*), *Tennis*, and *Tank*, though they are described in sufficient details in both languages, there tends to be less overlap among the text. The methods tend to retrieve more accurate similar pairs from person pages than other pages especially those pages describing a more abstract concepts. However, this needs to be tested more thoroughly.

When we look at the total number of sentence pairs returned, we notice that the bilingual lexicon based method consistently returns a smaller amount of similar sentence pairs which makes the method more accurate than the MT based approach. On average, the MT based approach returns 4.5 (26%) correct sentences and the bilingual lexicon based approach returns 2.9 correct sentences (45%). But, on average, the MT approach returns three times as many sentence pairs as bilingual lexicon approach. This may be due to the fact that the former makes use of restricted set of important terms or concepts whereas the later uses a large general lexicon. Though we remove some of the most frequently occuring stopwords in the MT based approach, it still generates a large number of incorrect similar sentence pairs due to some common words.

In general, the number of correctly identified similar pages extracted seems small. However, most of the Dutch pages are relatively small, which sets the upper bound on the number of correctly identified sentence pairs that can be extracted. On average, each Dutch Wikipedia page in the sample contains 18 sentences whereas English Wikipedia pages contain 65 sentences. Excluding the pages for *Tennis*, *Tank* (Dutch: *voertuig*), and *Tricolor*, which are relatively large, each Dutch page contains on average 8 sentences, which is even smaller. Given the fact that the pages are in general not parallel, the methods, using simple heuristics, identified high quality translation equivalent sentence pairs from most Wikipedia pages. Furthermore, a close examination of the output of the two approaches show that both tend to identify the same set of similar sentence pairs.

We ran our bilingual lexicon based approach on the whole Dutch-English Wikipedia corpus. The method returned about 80M of candidate similar sentences. Though we do not have the resources to evaluate this output, the results we got from sample data (cf. Table 1) suggest that it contains a significant amount of correctly identified similar

Title		MT		Bilingual Lexicon	
English	Dutch	Total	Match	Total	Match
Hersfeld Rotenburg	Hersfeld Rotenburg	2	_	3	2
Manganese nodule	Mangaanknol	5	2	1	1
Kettle	Ketel	_	_	1	1
Treason	Landverraad	2	_	1	_
Pierluigi Collina	Pierluigi Collina	14	13	13	11
Province of Ferrara	Ferrara (provincie)	7	1	1	1
Classicism	Classicisme	8	_	1	_
Tennis	Tennis	93	4	15	3
Hysteria	Hysterie	14	6	9	5
George F. Kennan	George Kennan	27	12	29	11
Marcus Cornelius Fronto	Marcus Cornelius Fronto	11	9	5	5
Delphi	Delphi (Griekenland)	34	2	8	1
De Beers	De Beers	11	5	10	5
Pavel Popovich	Pavel Popovytsj	7	4	4	4
Rice pudding	Rijstebrij	11	1	4	_
Manta ray	Reuzenmanta	15	3	7	2
Michelstadt	Michelstadt	1	1	1	1
Tank	Tank (voertuig)	84	3	27	2
Cheyenne(Wyoming)	Cheyenne(Wyoming)	5	2	2	2
Goa	Goa(deelstaat)	13	4	6	1
Tricolour	Driekleur	57	36	13	12
Oral cancer	Mondkanker	25	2	7	2
Pallium	Pallium	12	2	5	4
Ajanta	Ajanta	3	3	2	2
Captain Jack (band)	Captain Jack	16	3	2	2
Proboscis Monkey	Neusaap	15	6	4	1
Patti Smith	Patti Smith	6	2	4	2
Flores Island, Portugal	Flores (Azoren)	3	2	1	1
Mercury 8	Mercury MA 8	11	3	4	1
Mutation	Mutatie	16	4	6	3
Average		17.6	4.5	6.5	2.9

Table 1: Test topics (column 1 and 2). The total number of sentence pairs (column 3) and the number of correctly identified similar sentence pairs (column 4) returned by the MT based approach. The total number of sentence pairs (column 5) and the number of correctly identified similar sentence pairs (column 6) returned by the method using a bilingual lexicon.

sentences.

6 Conclusion

In this paper we focused on multilingual aspects of Wikipedia. Particularly, we investigated the potential of Wikipedia for generating parallel corpora by applying different methods for identifying similar text across multiple languages. We presented two methods and carried out an evaluation on a sample of Dutch-English Wikipedia pages. The results show that both methods, using simple heuristics, were able to identify similar text between the pair of Wikipedia pages though they differ in accuracy.

The bilingual lexicon approach returns fewer incorrect pairs than the MT based approach. We interpret this as saying that our bilingual lexicon based method provides a more accurate representation of the aboutness of sentences in Wikipedia than the MT based approach. Furthermore, the result we obtained on a sample of Wikipedia pages and the output of running the bilingual based approach on the whole Dutch-English gives some indication of the potential of Wikipedia for generating parallel corpora.

As to future work, the sentence similarity detection methods that we considered are not perfect. E.g., the MT based approach relies on rough translations; it is important to investigate the contribution of high quality translations. The bilingual lexicon approach uses only lexical features; other language specific sentence features might help improve results.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006, 640.-001.501, and 640.002.501.

References

D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. 2005. Using Wikipedia at the TREC QA Track. In E.M. Voorhees and L.P. Buckland, editors, *The Thirteenth Text Retrieval Conference (TREC 2004)*.

F. Bellomi and R. Bonato. 2005. Lexical authorities in an encyclopedic corpus: a case study with wikipedia. URL: `http://www.fran.it/blog/2005/01/lexical-authorities-in-encyclopedic.htm%l`. Site accessed on June 9, 2005.

A. Ciffolilli. 2003. Phantom authority, selfselective recruitment and retention of members in virtual communities: The case of Wikipedia. *First Monday*, 8(12).

S. Fissaha Adafre and M. de Rijke. 2005. Discovering missing links in Wikipedia. In *Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)*.

R. Ghani, S. Slattery, and Y. Yang. 2001. Hypertext categorization using hyperlink patterns and meta data. In Carla Brodley and Andrea Danyluk, editors, *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 178–185.

D. Kirk Evans. 2005. Identifying similarity in text: Multi-lingual analysis for summarization. URL: `http://www1.cs.columbia.edu/nlp/theses/dave_evans.pdf`. Site accessed on January 5, 2006.

A. Lih. 2004. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism*.

D. Melamed. 1996. A geometric approach to mapping bitext correspondence. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1–12, Somerset, New Jersey. Association for Computational Linguistics.

N. Miller. 2005. Wikipedia and the disappearing "Author". *ETC: A Review of General Semantics*, 62(1):37–40.

U. Rao and M. Turoff. 1990. Hypertext functionality: A theoretical framework. *International Journal of Human-Computer Interaction*.

F. Viégas, M. Wattenberg, and D. Kushal. 2004. Studying cooperation and conflict between authors with history flow visualization. In *Proceedings of the 2004 conference on Human factors in computing systems*.

J. Voss. 2005. Measuring Wikipedia. In *Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics*.

WiQA. 2006. Question answering using Wikipedia. URL: `http://ilps.science.uva.nl/WiQA/`. Site accessed on January 5, 2006.