

The Wikipedia Corpus

Jeff Pasternack and Dan Roth

Department of Computer Science
University of Illinois, Urbana-Champaign
{jpaster2,danr}@uiuc.edu

Abstract

Wikipedia, the popular online encyclopedia, has in just six years grown from an adjunct to the now-defunct Nupedia to over 31 million pages and 429 million revisions in 256 languages and spawned sister projects such as Wiktionary and Wikisource. Available under the GNU Free Documentation License, it is an extraordinarily large corpus with broad scope and constant updates. Its articles are largely consistent in structure and organized into category hierarchies. However, the wiki method of collaborative editing creates challenges that must be addressed. Wikipedia's accuracy is frequently questioned, and systemic bias means that quality and coverage are uneven, while even the variety of English dialects juxtaposed can sabotage the unwary with differences in semantics, diction and spelling. This paper examines Wikipedia from a research perspective, providing basic background knowledge and an understanding of its strengths and weaknesses. We also solve a technical challenge posed by the enormity of text (1.04TB for the English version) made available with a simple, easily-implemented dictionary compression algorithm that permits time-efficient random access to the data with a twenty-eight-fold reduction in size.

Introduction

Wikipedia has become one of the most frequented destinations on the web with, as of January 2008, a three-month average popularity ranking of 9th with 8.46% of all Internet users visiting it on a given day as estimated by Alexa Internet, up from 12th and 5.74% a year ago. Collaboratively authored and updated by its readers, the English version has more than 3.4 million registered users and “anonymous” edits by more than 4.5 million unique IP addresses. There are 796,264 non-stub encyclopedia articles with an average of 5,113.4 characters each in the English version, including markup¹, with another 618,237 short proto-articles averaging 1,405.9 characters that are either explicitly marked as stubs or lack links to other pages. Most articles feature metadata: on average a non-stub article belongs to 2.76 categories and has 2.94

template references (which function as rich, parameterized tags). Wikipedia is also highly responsive to current events: when Steve Irwin died on September 4th, 2006, his biography was edited 1,790 times that day alone.

However, as one may expect, there are dangers in employing an ever-changing resource with millions of amateur, faceless contributors (a quarter of all edits are anonymous), and in this paper we seek to delineate these so that the bias they introduce may be avoided, overcome, or at least recognized. While this task is usually easy for Wikipedia's human readers, a naïve algorithm will likely stumble without even realizing it. Consider, for example, a simple function that evaluates the similarity of two topics by calculating the overlap between their bags of nouns. Because of varying dialects used by contributors (only 54.9% of anonymous edits were from the United States) it will fail to match “soccer” and “football” but will mistakenly match “cot” where one editor meant “crib” and another meant “collapsible bed”. Software building a knowledge base from the corpus, on the other hand, might be stymied by the multitude of contradictions or even simple malicious page-blanking, lacking the sophistication to refer to previous versions of an article or to note the metadata indicating disputed factual accuracy or frequent vandalism. It might also be misled by ambiguous statistics (Hong Kong dollars or Australian dollars?) or assume nation-specific information to be universally true—a liberal in the U.S. is not the same as a liberal in the U.K., and “overseas” is by itself a meaningless location (contributors are often incognizant that other readers do not share their frame of reference). Similarly, it is hazardous to assume that a topic's real-world importance correlates with its article's size or number of citations in other articles: Wikipedia editors are self-selected and tend to favor topics of personal interest, and what is “notable” enough to be included is highly subjective (and frequently contentious).

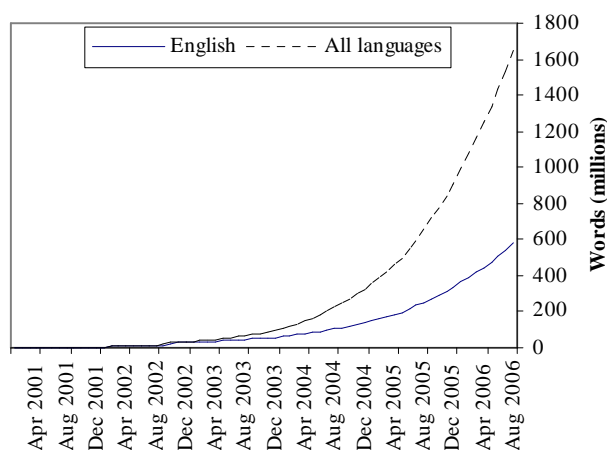
As we shall see, while these faults do require caution, there remains opportunity to exploit a rapidly expanding body of knowledge and semi-structured, annotated text for tasks in natural language processing, information retrieval and filtering, and even vision (many articles include template-tagged images), among other fields. The corpus also includes both current and past versions of each page, providing a complete provenance for an article that has

¹ Unless otherwise specified, all statistics presented in this paper are based on the English Wikipedia as of 11/4/2006, the most recent complete dump available to the authors at the time of this writing.

been mostly ignored by research to date. This may partly be due to the technical difficulty of manipulating over a terabyte of text. Our dictionary compression algorithm, as discussed later, solves this by both greatly reducing the size of the text as well as permitting fast random access to corpus pages and their past revisions.

A Brief History of Wikipedia

Wikipedia was founded in January 2001 by Jimmy Wales as a community-authored complement to Nupedia, which was a more traditional expert-written, peer-reviewed encyclopedia. Nupedia never realized more than twenty-four articles and was finally shut down in September 2003 (Wikipedia, 2008a); Wikipedia, on the other hand, was a success—the English version alone grew from roughly 43 thousand words in February 2001 to 13.3 million words in September 2002. In the next month the word count doubled to 26.2 million, though this increase was almost entirely the result of automatically creating roughly 36 thousand articles about towns in the United States based on US census information (Lih, 2004), (Wikipedia, 2008b). From October 2002 through 2006, the English version approximately doubled its word count every year, and Wikipedia on the whole grew even faster². In 2007, the number of pages in the English Wikipedia almost doubled, from six million to eleven million, but the number of articles (as defined by Wikimedia) increased only 38%, from 1.6 to 2.2 million, compared to 74% growth in 2006 and 104% in 2005, suggesting that new article creation is slowing in favor of additional auxiliary documents such as redirects and talk pages.



How Wikipedia Works

Wikipedia and its myriad sister projects (Wikionary, Wikibooks, Wikinews, Wikisource, Wikimedia Commons,

Wikiquote, Wikispecies, Wikiversity, and Meta) are operated by the Wikimedia Foundation, a non-profit corporation created in 2003, and all run on MediaWiki software (which is also used for many other wikis not affiliated with the Foundation). MediaWiki allows users to “collaboratively author” a website’s content by editing its constituent pages in wikitext, a specialized markup language. Each edit creates a revision, a new, altered copy of the page. A page’s revision history is thus the sequence of versions of that page from when it was first created until its most recent evolution. When a visitor requests a page, they are shown the latest revision, rendered from wikitext into HTML by the MediaWiki server. Some of the most contentious and popular Wikipedia articles such as “George W. Bush” have tens of thousands of revisions, but the average number of revisions per page is only 12.8. As will be discussed, however, most Wikipedia pages are not encyclopedia articles and the number of revisions per article is substantially higher, up to 56.6, depending on where the boundary is drawn between articles and non-articles. Some (typically new) pages judged to be inappropriate by the Wikipedia community may be deleted outright, but otherwise a page’s revisions (including vandalism) are kept forever and accessible at any time, the exceptions being those that create liability (e.g. copyright infringement or libel) for the Wikimedia Foundation.

Wikipedia does not have a formal process of peer review, and articles can be edited by any visitor, not just subject matter experts. Consequently, despite having a “neutral point of view” policy (Wikipedia, 2007c), disputes often arise. Although Wikipedia contains aspects of many forms of governance (Wikimedia, 2008), the contents of a page are typically decided by a mix of democracy and anarchy—contentious changes are sometimes discussed on a related “talk” page (e.g. Talk:George W. Bush) and subjected to a vote, while at other times implemented by an editor’s unilateral action, frequently leading to “edit wars” (Wikipedia, 2008d), where users undo each others’ alterations. The quality and completeness of an article, therefore, does not necessarily increase monotonically.

This, combined with recent controversy (most notably the libelous biography of John Seigenthaler (Seigenthaler, 2005)), has cast public doubt upon Wikipedia’s reliability. Nature’s well-known study (Giles, 2005), however, found that “the difference in accuracy was not particularly great” when comparing sampled scientific articles to those of the more traditional Encyclopaedia Britannica. There has also been a drive to provide additional supporting references and check contributions for validity. Still, errors are fairly common: as of January 2008, 2459 articles were formally marked as having disputed factual accuracy, and far more inaccuracies either go unnoticed (especially on less popular articles) or are disputed informally, usually on the article’s “talk” page. While researchers have long dealt with noisy data and Wikipedia is generally considered to be “by-and-large” reliable by its human audience, it is nevertheless unsafe to take anything presented within at face value. Some work, (McGuinness et al., 2006) and (Adler et al.,

² Detailed statistics can be found at <http://stats.wikimedia.org/EN/Sitemap.htm>

2007), has already begun considering “trust” in wikis which may provide a more systematic approach to the problem.

Corpus Overview

Wikipedia consists of 256 different language-specific versions. The English version is the largest with approximately 35% of all words. Each page in the corpus belongs to one of eighteen namespaces (table 1). Namespaces are prepended to page names, so the category page for algorithms is “Category:Algorithms”. If no namespace is specified for a page (e.g. “George W. Bush”) then the default namespace is assumed. The nine talk namespaces serve as message boards to facilitate discussion of pages and topics of the other nine, so “User talk” pages discuss Wikipedia users and their personal pages, and “Talk” pages discuss encyclopedia articles.

Namespace	Description (Example)
User (User talk)	Personal pages for and about Wikipedia users (User:Jimbo Wales)
Wikipedia (Wikipedia talk)	Metainformation about Wikipedia use, administration and editing (Wikipedia:About)
Image (Image talk)	Descriptions of image or sound files (Image:Osama-med.jpg)
Template (Template talk)	Pages that can be embedded in other pages, similar to Server Side Includes (Template:Disambig)
Category (Category talk)	Information about categories (displayed along with an automatically generated list of pages in that category) (Category:People)
[Default] (Talk)	Encyclopedia articles, redirects, disambiguation pages, and article stubs (Sabbatai Zevi)

Table 1: Important Wikipedia namespaces; talk namespaces and examples are in parentheses

A page’s membership in a category is determined simply by a wikitext link to it, so the article “Greedy algorithm” indicates its membership in the “Algorithms” category by including the wikitext link “[[Category:Algorithms]]”. Nearly all articles belong to at least one category. Since category information is decentralized and category listings are generated automatically, category pages themselves do not include lists of their constituents but rather link to other categories to describe a category hierarchy (which may contain cycles, e.g. “Category:Health” is a subcategory of “Category:Medicine”, and vice versa).

Like categories, templates can also provide semantic metadata about a topic. The Persondata template, for example, tags biographical articles with machine-readable

information on their subjects. Alternatively, a template can indicate something about the page itself, such as the NPOV template for pages with disputed neutrality. Double curly braces are used to reference the template with any required pipe-delimited arguments in the page’s wikitext (e.g. “{{Persondata|NAME=...}}”). The Mediawiki software replaces these references with the wikitext of the template when the page is rendered. Otherwise, however, the actual templates are not useful: template references alone supply the relevant metadata.

Finally, many popular articles link to the same topic in other languages (displayed in the languages sidebar); the English DNA article, for instance, uses the wikitext “[[nl:DNA]]” to point to the Dutch version and “[[de:Desoxyribonukleinsäure]]” for German. Wikipedia is thus a very sizable potential resource for machine translation, although it could be significantly more difficult to exploit than a traditional multilingual corpus as each language’s version is independently created and edited, and structure, coverage and depth may vary widely.

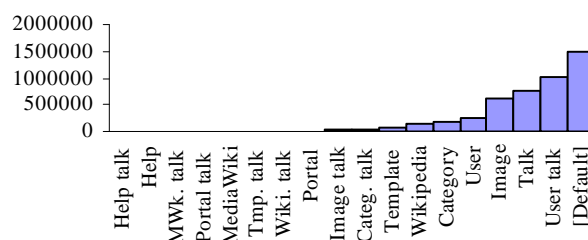


Figure 2: Page counts by namespace, excluding redirects (4,662,123 pages total)

Articles and Non-articles

Not all pages in the default namespace are encyclopedia articles, and classification as such can be subjective according to one’s needs. Wikipedia itself defines articles as belonging to the default namespace but excludes pages that are “redirects”, have no links to other Wikipedia pages or serve only to disambiguate a term (Wikipedia, 2008e). Redirects have no content but rather direct a user to another page; “George Bush Jr.”, for example, redirects to “George W. Bush”. A disambiguation page, on the other hand, links to a term’s possible meanings: “George Bush” lists links to the 41st and 43rd US presidents as well as other notable people and things with that name (such as the George Bush Intercontinental Airport). Both page types supply synonyms for a term (Bunescu and Pasca, 2006).

Of the 2,954,265 pages in the default namespace, almost half (1,456,736) are redirects, and 76,472 are disambiguation pages. If we also set aside the 13,696 remaining pages with no links to other Wikipedia pages, we are left with a total of 1,407,374 articles with an average 38.4 revisions each. This figure is misleading, however, as it does not account for article “stubs” (Wikipedia, 2008f), very short pages that may consist of no more than a sentence, having 1,403.1 characters of wikitext on average compared to 5,113.4 characters for non-stub

articles. Discounting the 611,110 stubs with internal links leaves 796,264 articles with 56.6 revisions each. Both stubs and disambiguation pages are marked by template references and are thus easy to recognize.

Advantages and Dangers

(Toral and Munoz, 2006) list some of the advantages of the Wikipedia corpus:

- Its size, over 18 million pages and 1.6 billion words.
- Content is made available under the GNU Free Documentation License, permitting free use by researchers.
- As an encyclopedia, its information is broad in scope.
- Pages have metadata indicated by categories and templates in addition to a somewhat regular structure.
- Multiple languages allow for non-English and multilingual applications (e.g. translation or cross-language named entity mapping).
- Content is ever-evolving and constantly updated.

The problems that arise from creating documents in the decentralized, chaotic “wiki way”, though, are not easily solved (Denning et al., 2005). Multiple editors make articles prone to inconsistency, and it is not unusual to find an article contradicting itself or other articles. Because Wikipedia attracts users worldwide the English dialect used varies between and sometimes within pages with no standardization: the British English word “petrol”, for example, appears 5,337 times compared to 8,788 appearances of the American English equivalent “gasoline”. Besides differing diction, spelling differences are also endemic—“organisation” appears 70,895 times versus 238,821 times for “organization” (including plurals). While these incongruities are of broad concern and relatively easy to detect and correct for, divergent semantics are a more subtle trap for tasks such as knowledge extraction or synset assignment. Units of measurement are often ambiguous (the U.S. gallon versus the imperial gallon or the U.S. dollar versus the Australian dollar) as are some dates (1/2/2007 may mean January 2nd or February 1st) and many words have divergent meanings, e.g. corn (British: grain, American: maize) or entrée (British: appetizer, American: main course).

Wikipedia’s general policy of allowing anyone to edit articles has, as already mentioned, resulted in much controversy due to the oft-realized potential for factual inaccuracy. Although obvious vandalism, such as deleting all article text, is usually quickly reverted (Viegas, Wattenberg, and Dave, 2004), the mean time between a revision explicitly labeled as reverting vandalism with “rvv” and the previous revision is 11.86 hours. In the wake of this, libel fears have resulted in shorter, less complete biographies of living persons. While in principle these are merely limited to what can be verified with reliable references (Wikipedia, 2008g), in practice many verifiable but unflattering facts are omitted. In other articles, controversial statements may be presented from several points of view, or they may similarly be eliminated

entirely (Stvilia et al., 2005). Other systemic biases can largely be attributed to the user base: Internet-savvy people with the time and resolve to contribute their time to an online encyclopedia and, in the English version, people predominantly from developed Commonwealth nations and the United States. Articles in fields such as technology and current events tend to be more numerous, longer, and higher in quality than less favored realms—the article for the video game character Sonic the Hedgehog (66,680 characters) is more than seven times as long as that for Indian art (9,394 characters), for example. Ignoring these biases will negatively affect results, e.g. a search algorithm might incorrectly conclude that Sonic is more likely to be of interest than Indian art because of the greater quantity of citations to it from other articles and websites, when in fact this merely because Sonic aficionados are more prolific contributors than the larger majority far more interested in the culture of India.

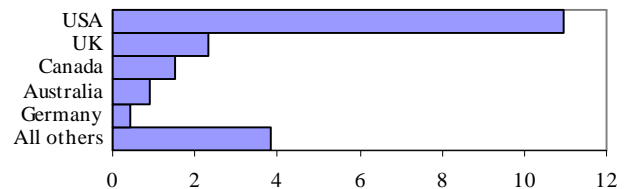


Figure 3: Top 5 countries by millions of anonymous edits

Research with Wikipedia

Existing research on Wikipedia can be broken into three groups: research on improving Wikipedia, research about Wikipedia, and research using Wikipedia as a corpus.

For those interested in Wikipedia as a corpus, research on improving Wikipedia is unlikely to be of immediate interest but rather suggests the direction it may take in the future. (Krotzsch, Vrandečić, and Volkel, 2005) have examined the use of “typed links” to convey semantic information, e.g. a link from Microsoft to Corporation would be typed as an “is-a” relationship. (Adafre and Rijke, 2005), on the other hand, attempt to find new inter-article links using clustering and heuristics to identify appropriate anchor text and targets. Lastly, (McGuinness et al., 2006) suggest a method to judge the reliability of content using a combination of provenance annotation and link structure, while (Adler et al., 2007) trusts contributors whose edits survive other editors.

Research about Wikipedia is often concerned with the social phenomenon of collaborative document authorship or the information quality that results. (Stvilia et al., 2005) considers in depth how users identify, debate and resolve these quality issues by studying “talk” pages and analyzing the tradeoffs (e.g. completeness vs. brevity). (Kittur et al., 2007) studies the costs of this community overhead further, demonstrating that, as Wikipedia grows, work maintaining and validating content (reverting vandalism, resolving disputes, etc.) is intensifying at the expense of content creation. (Lih, 2004) examines the effect that citation of

Wikipedia articles by the press has on their quality, as measured by the number of edits and unique editors, finding (unsurprisingly) that the additional traffic increased these counts. (Viegas, Wattenberg, and Dave, 2004) studied collaborative authorship in Wikipedia with “history flows”, visualizations of an article’s revision history. Finally, (Voss, 2005) presents a summary of Wikipedia with a number of detailed statistics and graphs, while (Voss, 2006) more narrowly focuses on its category hierarchies as a thesaurus, comparing it other classification systems such as Dewey Decimal.

The tasks to which the Wikipedia corpus has been applied thus far often rely on a small, well-structured subset of the data such as categorization or interarticle links. Most work ignores the pitfalls we have discussed, suggesting that the mediocre results some have achieved could be improved by even simple measures such as accounting for spelling differences among English dialects. (Ruiz-Casado, Alfonseca, and Castells, 2005a) use WordNet to find an article concept’s synset (Ruiz-Casado, Alfonseca and Castells, 2005b) and next collects article sentences containing hyperlinks that reflect known WordNet relationships, which are then employed to identify new relations. (Ponzetto and Strube, 2006) also uses WordNet and Wikipedia in addition to the ASSERT semantic role labeler to create a coreference resolution system. (Strube and Ponzetto, 2006), (Gabrilovich and Markovitch, 2007) and (Milne, 2007) all utilize Wikipedia to compute semantic relatedness. (Toral and Munoz, 2006) propose the automatic extraction of gazetteers for named entity recognition from Wikipedia while (Bunescu and Pasca, 2006) use the corpus to learn to disambiguate named entities in the context of web search. Finally, (Adafre and Rijke, 2006) attempt to synthesize multilingual parallel corpora by applying simple heuristics to match corresponding sentences across different Wikipedia language versions.

Technical Considerations

The most recent page revisions are published every month or two at <http://download.wikipedia.org> as compressed XML files. Complete revision histories of the English corpus, however, are available only sporadically as these dumps frequently fail due to their size.

Dictionary Compression for Revision Histories

Relatively little research to date has exploited the full revision history that is available for each article, instead preferring the single most-recent version. Researchers have rarely dealt with detailed document provenance in the past and so may be oblivious to the opportunities it affords; one could, for example, track current events by correlating the time of a revision to the facts it introduces, identify controversial contributions by the ensuing edit war (which can then be discarded as noise), or build an ontology where the categorization of a concept is the first category to

appear in the article’s revision history (earlier categorizations are likely to be more fundamental: “The Jetsons” now has nineteen categories, but the first was “animation”). There is, however, a technical obstacle: the complete English version XML dump from November 2006 is 1.15TB (including 1.04TB of revision text) and will only continue to grow in the future. In the compressed form provided by Wikimedia (7zip or bzip2), though, these files allow only sequential access to pages and their revisions. This works well when one requires only a complete pass over the data (e.g. to gather simple statistics), but makes random access infeasible, requiring an algorithm to either keep all relevant portions of the corpus in RAM (often impossible) or page them to disk.

Storing the relevant data uncompressed on disk is sometimes viable, but once the data is written reading it back sequentially (32.9MB/s)³ is, perhaps counterintuitively, four times *slower* than 7zip (125.8MB/s), since the disk transfer rate is a far greater bottleneck than the CPU time required for decompression. Another alternative that allows random access to pages (though not revisions) is employed by the MediaWiki software, which stores only the differences between subsequent versions rather than the full text (e.g. as might be provided by the diff utility). Since most edits are small, this yields high compression and low I/O load but, conversely, requires a great deal of CPU time for both compression and decompression.

Our solution is to instead exploit knowledge of Wikipedia to implement a more suitable dictionary-based compression scheme that will allow random access to both pages and revisions. Revisions can be readily divided into segments by their newlines, providing an easy means of separating article paragraphs, references, headers and so on. There are, on average, 136.4 segments per revision, but only 2.84 of these newline-delimited segments are changed in each revision, suggesting the following compression algorithm:

```
//bijective map of segments to IDs, IDs to segments
Bijjective hashtable segmentTable
Integer nextID=0;
For each revision of the page being compressed:
  Split the revision text on newlines
  For each newline-delimited segment s:
    If s is in the segmentTable, write its ID to output
    Otherwise,
      Add (s, nextID) to the segmentTable
      Write nextID to output
      nextID = nextID + 1
Save the segmentTable for future decompression
```

A few bytes of metadata may be written to the output for each revision to specify the number of newline-delimited

³ All benchmarks are averaged data rates of sequential reads of the November 2006 corpus using single-threaded implementations on an Intel 2.66GHz Core Duo system with 2GB DDR2 800 RAM and 7200RPM hard disks.

segments and the number of bits used for each segment ID. Decompressing a revision is simple: read its list of segment IDs and then replace each of those with the corresponding text from the segment table. It is slightly faster than 7zip, at 133.2MB/s.

Most importantly, pages and revisions are available via random-access, since only the segments needed for the desired revisions must be read from the on-disk segment table. Revision text is compressed from 1062.4GB to 37.66GB and the total size including uncompressed metadata (e.g. page titles, editor comments and identities, and timestamps) is 58.0GB. Wikimedia's 7zip compressed XML dump, on the other hand, is just 8.1GB, so there is a tradeoff of space efficiency for speed and random access.

Code for a C# implementation of the dictionary compression algorithm, Wikipedia object model and an XML dump parser may be found at the author's website.

Conclusion

Wikipedia is a growing resource with substantial untapped potential and myriad benefits, but this is tempered by the uncertainties and challenges arising from the "wiki" method of collaborative authorship that can easily confound results. To facilitate the use of Wikipedia as a corpus we have thus provided an overview of its structure and metadata, and explored both its advantages and dangers in detail, demonstrating that the latter can be largely overcome with proper caution and sufficient domain knowledge. In particular, our specialized compression algorithm has solved the problem posed by the enormous size of complete revision histories and so opened the door to the novel possibilities they enable.

References

- Adafre, S., and de Rijke, M. 2005. Discovering missing links in Wikipedia. *LinkKDD-2005*.
- Adafre, S., and de Rijke, M. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. *Workshop on New Text, EACL 2006*.
- Adler, B.; Benterou, J.; Chatterjee, K.; de Alfaro, L.; Pye, I.; and Raman, V. 2007. Assigning Trust To Wikipedia Content. Technical Report UCSC-CRL-07-09, School of Engineering, University of California, Santa Cruz.
- Bunescu, R., and Pasca, M. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation. *EACL-06*.
- Denning, P.; Horning, J.; Parnas, D.; and Weinstein, L. 2005. Wikipedia risks. *Communications of the ACM* 48(12):152–152.
- Gabrilovich, E., and Markovitch, S. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *IJCAI 2007*.
- Giles, J. 2005. Internet encyclopedias go head to head. *Nature* 12/14/2005.
- Kittur, A.; Suh, B.; Pendleton, B.; and Chi, E. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. *SIGCHI 2007*.
- Krotzsch, M.; Vrandečić, D.; and Volkel, M. 2005. Wikipedia and the Semantic Web-The Missing Links. *Wikimania 2005*.
- Lih, A. 2003. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. *Nature* 2004.
- McGuinness, D.; Zeng, H.; da Silva, P.; Ding, L.; Narayanan, D.; and Bhaowal, M. 2006. Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study. *MTW 2006*.
- Milne, D. 2007. Computing Semantic Relatedness using Wikipedia Link Structure. *NZCSRSC 2007*.
- Ponzetto, S., and Strube, M. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. *HCT 2006*.
- Ruiz-Casado, M.; Alfonseca, E.; and Castells, P. 2005a. Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia. *NLDB 2006*.
- Ruiz-Casado, M.; Alfonseca, E.; and Castells, P. 2005b. Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. *AWIC 2005*.
- Seigenthaler, J. 2005. A false Wikipedia 'biography'. *USA Today* 11/29/2005.
- Strube, M., and Ponzetto, S. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. *AAAI06*.
- Stvilia, B.; Twidale, M.; Gasser, L.; and Smith, L. 2005. Information Quality Discussions in Wikipedia. *ICKM '05*.
- Toral, A., and Muñoz, R. 2006. A proposal to automatically build and maintain gazetteers for Named Entity Recognition using Wikipedia. *Workshop on New Text, EACL 2006*.
- Viégas, F.; Wattenberg, M.; and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations. *SIGCHI 2004*.
- Voss, J. 2005. Measuring Wikipedia. *ISSI 2005*.
- Voss, J. 2006. Collaborative thesaurus tagging the Wikipedia way. *ArXiv Computer Science e-prints*.
- Wikimedia. 2008. Power Structure. http://meta.wikimedia.org/wiki/Power_structure.
- Wikipedia. 2008a. Nupedia. *Wikipedia, The Free Encyclopedia*. <http://en.wikipedia.org>.
- Wikipedia. 2008b. History of Wikipedia.
- Wikipedia. 2008c. Wikipedia:Neutral point of view.
- Wikipedia. 2008d. Wikipedia:Edit War.
- Wikipedia. 2008e. Wikipedia:What is an article.
- Wikipedia. 2008f. Wikipedia:Stub.
- Wikipedia. 2008g. Wikipedia:Biographies of living persons.