Working Paper Series
ISSN 1177-777X

MINING MEANING FROM WIKIPEDIA

**Olena Medelyan, Catherine Legg,
David Milne and Ian H. Witten**

Working Paper: 11/2008
September 2008

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Mining meaning from Wikipedia

OLENA MEDELYAN, CATHERINE LEGG, DAVID MILNE and IAN H. WITTEN
University of Waikato, New Zealand

Wikipedia is a goldmine of information; not just for its many readers, but also for the growing community of researchers who recognize it as a resource of exceptional scale and utility. It represents a vast investment of manual effort and judgment: a huge, constantly evolving tapestry of concepts and relations that is being applied to a host of tasks.

This article provides a comprehensive description of this work. It focuses on research that extracts and makes use of the concepts, relations, facts and descriptions found in Wikipedia, and organizes the work into four broad categories: applying Wikipedia to *natural language processing*; using it to facilitate *information retrieval* and *information extraction*; and as a resource for *ontology building*. The article addresses how Wikipedia is being used as is, how it is being improved and adapted, and how it is being combined with other structures to create entirely new resources. We identify the research groups and individuals involved, and how their work has developed in the last few years. We provide a comprehensive list of the open-source software they have produced. We also discuss the implications of this work for the long-awaited semantic web.

1. INTRODUCTION

Wikipedia requires little introduction or explanation. As everyone knows, it was launched in 2001 with the goal of building free encyclopedias in all languages. Today it is easily the largest and most widely-used encyclopedia in existence. Wikipedia has become something of a phenomenon among computer scientists as well as the general public. It represents a vast investment of freely-given manual effort and judgment, and the last few years have seen a multitude of papers that apply it to a host of different problems. This paper provides the first comprehensive summary of this research (up to mid-2008), which we collect under the deliberately vague umbrella of *mining meaning from Wikipedia*. By *meaning*, we encompass everything from concepts, topics, and descriptions to facts, semantic relations, and ways of organizing information. *Mining* involves both gathering meaning into machine-readable structures (such as ontologies), and using it in areas like information retrieval and natural language processing.

Traditional approaches to mining meaning fall into two broad camps. On one side are carefully hand-crafted resources, such as thesauri and ontologies. These resources are generally of high quality, but by necessity are restricted in size and coverage. They rely on the input of experts, who cannot hope to keep abreast of the incalculable tide of new discoveries and topics that arise constantly. Even the most extensive manually created resource—the Cyc ontology, whose hundreds of contributors have toiled for 20 years—has limited size and patchy coverage [Sowa 2004]. The other option is to sacrifice quality for quantity and obtain knowledge by performing large-scale analysis of unstructured text. However, human language is rife with inconsistency, and our intuitive understanding of it

cannot be entirely replicated in rules or trends, no matter how much data they are based upon. Approaches based on statistical inference might emulate human intelligence for specific tasks and in specific situations, but cracks appear when generalizing or moving into new domains and tasks.

Wikipedia provides a middle ground between these two camps—quality and quantity—by offering a rare mix of scale and structure. With two million articles and thousands of contributors, it dwarfs any other manually created resource by an order of magnitude in the number of concepts covered, has far greater potential for growth, and offers a wealth of further useful structural features. It contains around 18 Gb of text, and its extensive network of links, categories and infoboxes provide a variety of explicitly defined semantics that other corpora lack. One must, however, keep Wikipedia in perspective. It does not always engender the same level of trust or expectations of quality as traditional resources, because its contributors are largely unknown and unqualified. It is also much smaller and less representative of all human language use than the web as a whole. Nevertheless, Wikipedia has received enthusiastic attention as a promising natural language and informational resource of unexpected quality and utility. Here we focus on research that makes use of Wikipedia, and as far as possible leave aside its controversial nature.

This paper is structured as follows. In the next section we describe Wikipedia's creation process and structure, and how it is viewed by computer scientists as anything from a corpus, taxonomy, thesaurus, or hierarchy of knowledge topics to a full-blown ontology. The next four sections describe different research applications. Section 3 explains how it is being drawn upon for *natural language processing*; understanding written text. In Section 4 we describe its applications for *information retrieval*; searching through documents, organizing them and answering questions. Section 5 focuses on *information extraction*; mining text for topics, relations and facts. Section 6 describes uses of Wikipedia for *ontology building*, and asks whether this adds up to Tim Berners-Lee's long-delayed vision of the semantic web. Section 7 documents the people and research groups involved, while Section 8 lists the resources they have produced, with URLs. The final section gives a brief overall summary.

2 WIKIPEDIA: A RESOURCE FOR MINING MEANING

Wikipedia, one of the most visited sites on the web, outstrips all other encyclopedias in size and coverage. Its English language articles alone are 10 times the size of the Encyclopedia Britannica, its nearest rival. But material in English constitutes only a quarter of Wikipedia—it has articles in 250 other languages as well. Co-founder Jimmy

Wales is on record as saying that he aspires to distribute a free encyclopedia to every person on the planet, in their own language.

This section provides a general overview of Wikipedia, as background to our discussions in Sections 3–6. We begin with an insight into its unique editing methods, their benefits and challenges (Section 2.1); and then outline its key structural features, such as articles, hyperlinks and categories (Section 2.2). In Section 2.3 we identify some different roles that Wikipedia as a whole may usefully be regarded as playing—for instance, as well as an encyclopedia it can be viewed as a linguistic corpus. We conclude in Section 2.4 with some practical information on how to work with Wikipedia data.

2.1 The Encyclopedic Wisdom of Crowds

From its inception the Wikipedia project offered a unique, entirely open, collaborative editing process, scaffolded by then-new wiki software for group website building, and it is fascinating to see how the resource has flourished under this system. It has effectively enabled the entire world to become a panel of experts, authors and reviewers—contributing under their own name, or, if they wish, anonymously.

In its early days the project attracted widespread skepticism. It was thought that its editing system was so anarchic that it would surely fill up with misconceptions, outright lies, vanity pieces and other worse-than-useless human output. A piece in The Onion satirical newspaper "Wikipedia Celebrates 750 Years Of American Independence: Founding Fathers, Patriots, Mr. T. Honored"[1] nicely captures this point of view. Moreover, it was argued, surely the ability for anyone to make any change, on any page, entirely anonymously, would leave the resource ludicrously vulnerable to vandalism, particularly to articles that cover sensitive topics. What if the hard work of 2000 people were erased by one eccentric? And indeed, "edit wars" did erupt, though it turned out that some of the most vicious raged over such apparently trivial topics as the ancestry of Freddy Mercury and the true spelling of yoghurt. Yet this turbulent experience was channeled into developing a set of ever-more sophisticated Wikipedia policies and guidelines,[2] as well as a more subtle code of recommended good manners referred to as Wikiquette.[3] A self-selecting set of administrators emerged, who performed regulatory functions such as blocking individuals from editing for periods of time—for instance edit warriors, identified by the fact that they "revert" an article more than three times in 24 hours. Interestingly, the development of these rules was guided by the goal of reaching consensus, just as the encyclopedia's content is.

---

[1] *http://www.theonion.com/content/node/50902*
[2] *http://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines*
[3] *http://en.wikipedia.org/wiki/Wikipedia:WQT*

Somehow these processes worked sufficiently to shepherd the resource through its growing pains, and today Wikipedia is wildly popular and growing all the time. Section 2.3.1 discusses its accuracy and trustworthiness as an encyclopedia.

There is still skepticism. For example, Magnus [2006], a philosopher, argues that Wikipedia does not enable him to use the methods he usually uses to "assess claims," such as relying on the reputation of the source, assessing whether the claims are written in an appropriate style or have content that sounds plausible to him. However, these observations can be placed in the context of larger philosophical discussions about the nature of knowledge and truth: potentially challenging contemporary philosophical wisdom itself. In many ways the history of the so-called "modern" period in Western culture—the 300 years or so since the Scientific Revolution—may be seen as the struggle to escape a medieval conception of knowledge as defined by some kind of stamp of approval conferred on human beliefs by a recognized authority. The key medieval authorities were the Bible and Aristotle, and although humanity now avails itself of many more sources of information, including scientific experiments, arguably Universities still claim the same kind of authoritative role as validators of knowledge, in particular through the peer review process, which underpins what is published. The received wisdom is that surely some external source or body has to validate knowledge claims, or where would we be? Yet Wikipedia threatens to tear this function from the academy. Many scholars have noticed this, and some fight back—for instance by banning students from using it [Baker 2008].

Other models of knowledge have been offered, however, that cast Wikipedia's success in a new light. In the late 19th century the pragmatist Peirce proposed that beliefs be understood as knowledge due not to their prior justification, but to their usefulness, public character and future development. His account of knowledge was based on a unique account of *truth*, which claimed that true beliefs are those that all sincere participants in a "community of inquiry" would converge on, given enough time. Influential 20th century philosophers [e.g. Quine 1960] scoffed at this notion as being insufficiently objective. Yet Peirce claimed that there is a kind of person whose greatest passion is to render the Universe intelligible and will freely give time to do so, and that over the long run, within a sufficiently broad community, the use of signs is intrinsically self-correcting [Peirce 1868]. Wikipedia can be seen as a fascinating and unanticipated concrete realization of these apparently wildly idealistic claims.

In this context it is interesting to note that Larry Sanger, Wikipedia co-founder and editor-in-chief, had his initial training as a philosopher—with a specialization in theory of knowledge. In public accounts of his work he has tried to bypass vexed philosophical

discussions of truth by claiming that Wikipedians are not seeking it but rather a neutral point of view.[4] But as the purpose of this is to support every reader being able to build their own opinion, it can be argued that somewhat paradoxically this is the fastest route to genuine consensus. Interestingly, however, he and the other co-founder Jimmy Wales eventually clashed over the issue of expert opinion's role in Wikipedia. Thus, in 2007 Sanger diverged to found a new public online encyclopedia Citizendium[5] in an attempt to "do better" than Wikipedia, apparently reasserting validation by external authority, e.g. academics. Interestingly, although it is early days, Citizendium seems to lack Wikipedia's popularity and momentum.

Wikipedia's unique editing methods, and the issues that surround them, have complex implications for mining. First, unlike a traditional corpus, it is constantly growing and changing, so results obtained at any given time can become stale. Some research strives to measure the degree of difference between Wikipedia versions over time (though this is only useful insofar as Wikipedia's rate of change is itself constant), and assess the impact on common research tasks [e.g. Ponzetto and Strube 2007a]. Second, how are projects that incorporate Wikipedia data to be evaluated? If Wikipedia editors are the only people in the world who have been enthusiastic enough to write up certain topics (for instance, details of TV program plots), how is one to determine 'ground truth' for evaluating applications that utilize this information? The third factor is more of an opportunity than a challenge. The awe-inspiring abundance of manual labor given freely to Wikipedia raises the possibility of a new kind of research project, which would consist in encouraging Wikipedians themselves to perform certain tasks on the researchers' behalf (possibly tasks of a scale the researchers themselves could not hope to achieve). As we will see (for instance in Section 6), some have begun to glimpse this possibility, while others continue to view Wikipedia in more traditional "product" rather than "process" terms. At any rate, this research area sits on a fascinating interface between software and social engineering.

2.2. Wikipedia's structure

Traditional paper encyclopedias consist of articles arranged alphabetically, with internal cross-references to other relevant places in the encyclopedia, external references to the academic literature, and some kind of general index of topics. These structural features have been adapted by Wikipedia for the online environment, and some new features arising from the Wiki editing process have been added. The statistics presented in this section were obtained from a version of English Wikipedia released in July 2008.

---

[4] *http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view*
[5] *http://en.citizendium.org*

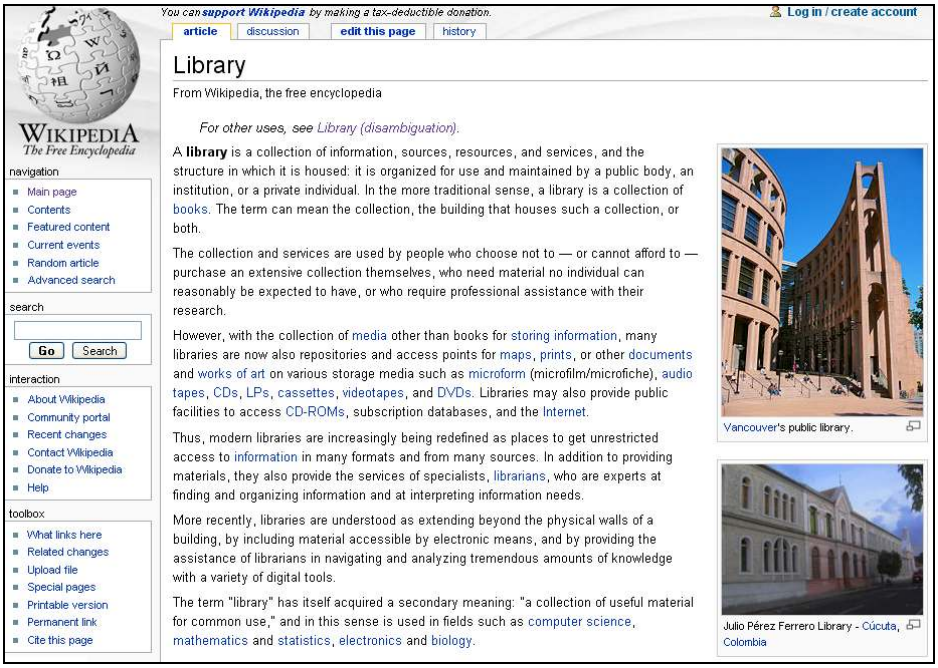

Figure 1. Wikipedia article on *Library*.

2.2.1. Articles: The basic unit of information in Wikipedia is the *article*. Internationally, Wikipedia contains 10M articles in its 250 different languages.[6] The English version contains 2.4M articles (not counting redirects and disambiguation pages, which are discussed below). About 1.8M of these are bona fide articles with more than 30 words of descriptive text and at least one incoming link from elsewhere in Wikipedia. Articles are written in a form of free text that follows a comprehensive set of editorial and structural guidelines in order to promote consistency and cohesion. These are laid down in the Manual of Style,[7] and include the following:

1. Each article describes a single concept, and there is a single article for each concept.
2. Article titles are succinct phrases that resemble terms in a conventional thesaurus.
3. Equivalent terms are linked to an article using redirects (Section 2.2.2).
4. Disambiguation pages present various possible meanings from which users can select an intended article. (Section 2.2.3).
5. Articles begin with a brief overview of the topic, and the first sentence defines the entity and its type.

[6] *http://en.wikipedia.org/wiki/Wikipedia*
[7] *http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style*

6. Articles contain hyperlinks that express relationships to other articles (Section 2.2.6).

Figure 1 shows a typical article, entitled *Library*. The first sentence describes the concept:

> A **library** is a collection of information, sources, resources, and services: it is organized for use and maintained by a public body, an institution, or a private individual.

Here the article's title is the single word *Library*, but titles are often qualified by appending parenthetical expressions. For example, there are other articles entitled *Library* (*computing*), *Library* (*electronics*), and *Library* (*biology*). Wikipedia distinguishes capitalization when it is relevant: the article *Optic nerve* (the nerve) is distinguished from *Optic Nerve* (the comic book).

2.2.2. Redirects: A redirect page is one with no text other than a directive in the form of a *redirect* link. There are about a dozen for *Library* and just under three million in the entire English Wikipedia; they encode pluralism (*libraries*), technical terms (*bibliotheca*), common misspellings (*libary*), and other variants (*reading room, book stack*). The aim is to have a single article for each concept and define redirects to link equivalent terms to the article's preferred title. As we will see, this helps with mining because to resolve synonymy an external thesaurus is unnecessary.

2.2.3. Disambiguation pages: Instead of taking readers to an article named by the term, as *Library* does, the Wikipedia search engine sometimes takes them directly to a special disambiguation page where they can click on the meaning they want. These pages are identified by invoking certain templates (discussed in Section 2.2.8) or assigning them to certain categories (Section 2.2.6), and often contain (*disambiguation*) in their title.

The English Wikipedia contains 100,000 disambiguation pages. The first line of the *Library* article in Figure 1 ("For other uses …") links to a disambiguation page that lists *Library* (*computing*), *Library* (*electronics*), *Library* (*biology*), and other senses of the term. Brief scope notes accompany each sense, to help users identify the correct one. For instance *Library* (*computer science*) is "a collection of subprograms used to develop software." The articles themselves serve as detailed scope notes. Disambiguation pages are helpful sources of information concerning homonyms.

2.2.5. Hyperlinks: Articles are peppered with hyperlinks to other articles: on average, about 25 of them. The English Wikipedia contains 60 million in total. They provide explanations of the topics being discussed and support an environment where serendipitous encounters with information are commonplace. Anyone who has browsed

Wikipedia has likely experienced the feeling of being happily lost, browsing from one interesting topic to the next and encountering information that they would never have searched for.

Wikipedia's hyperlinks are also useful from a linguistic standpoint. They are an additional source of synonyms that are not captured by redirects, because the terms used as anchors are often couched in different words. *Library*, for example, is referenced by 20 different anchors including *library*, *libraries*, and *biblioteca*. They also complement disambiguation pages by encoding polysemy; *library* links to different articles depending on the context in which it is found. They also give a sense of how well known each sense is; 84% of *library* links go to the article shown in Figure 1, while only 13% go to *Library* (*computing*). Furthermore, since hyperlinks in Wikipedia indicate that one article relates to another in some respect, this fundamental structure can be mined for meaning in many interesting ways—capturing the associative relations included in standard thesauri (Section 5.2), to give just one example.

2.2.6. Category structure: Authors are encouraged to assign categories to their articles. For example, the article *Library* falls in the category *Book Promotion*. Authors are also encouraged to assign the categories themselves to other more general categories; *Book Promotion* belongs to *Books*, which in turn belongs to *Written Communication*. These categorizations, like the articles themselves, can be modified by anyone. There are almost 400,000 categories in the English Wikipedia, with an average of 19 articles and two subcategories each.

Categories are not themselves articles. They are merely nodes for organizing the articles they contain, with a minimum of explanatory text. Often (in about a third of cases), categories correspond to a concept that requires further description. In these cases they are paired with an article of the same name: the category *Libraries* is paired with the article *Library*, and *Billionaires* with *Billionaire*. Other categories, such as *Libraries by country*, have no corresponding articles and serve only to organize the content. For clarity, in this paper we indicate categories in the form *Category:Books* unless it is obvious that we are not talking about an article.

The goal of the category structure is to represent information hierarchy. It is not a simple tree-structured taxonomy, but a graph in which multiple organization schemes coexist. Thus both articles and categories can belong to more than one category. The category *Libraries* belongs to four: *Buildings and structures*, *Civil services*, *Culture* and *Library and information science*. The overall structure approximates an acyclic directed graph; all relations are directional, and although cycles sometimes occur, they are uncommon. According to Wikipedia's own guidelines, cycles are generally discouraged

but may be acceptable in rare cases. For example, *Education* is a field within *Social Sciences*, which is an *Academic discipline*, which belongs under *Education*. In other words, you can educate people about how to educate.

A relatively recent addition to the encyclopedia, and less visible than articles, the category structure is haphazard, redundant, incomplete, and inconsistent [Chernov *et al.* 2006; Muchnik *et al.* 2007]. Links represent a wide variety of types and strengths of relationships. Although there has been much cleanup and the greatest proportion of links now represent class membership (*isa*), there are still many representing physical parthood, geographical location and many other merely thematic associations between entities—as well as meta-categories used for editorial purposes, such as *Disambiguation*. Thus *Category:Pork* currently contains, among others, the categories *Domestic Pig, Bacon Bits, Religious Restrictions on the Consumption of Pork*, and *Full Breakfast*. We will see in Section 6 that there are opportunities for recruiting users to help with data cleaning. We will also see in Section 5 that the issues mentioned above have not prevented researchers from innovatively and fruitfully mining the category structure for a range of different purposes.

2.2.8 Templates and infoboxes: Templates are pages that are not used in isolation, but are instead invoked to add information to other pages in a reusable fashion. Wikipedia contains 174,000 different templates, which have been invoked 23 million times. They are commonly used to identify articles that require attention; e.g. if they are biased, poorly written, or lacking citations. They can also define pages of different types, such as disambiguation pages or featured (high quality) articles. A common application is to provide navigational links, such as the *for other uses* link shown in Figure 1.

An infobox is a special type of template that displays factual information in a structured uniform format. Figure 2 shows one from the article on the *Library of Congress*. It was created by invoking the *Infobox Library* template and populating its fields, such as *location* and *collection size*. There are 8,000 different infobox templates that are used for anything from animal species to strategies for starting a game of chess, and the number is growing rapidly.

There are several simple ways in which the infobox structure could be improved. Standard representations for units would allow quantities to be extracted reliably. Different attribute names are often used for the same kind of content. More far-reaching would be to associate data types with attribute values, and allow language and unit tags when information can be expressed in different ways (e.g. Euro and USD). Many Wikipedia articles use tables for structured information that would be better represented as templates [Auer and Lehmann 2007]. Despite these problems, it is surprising how much

meaningful and machine-interpretable information can be extracted from Wikipedia templates. This is discussed further in Sections 5.3 and 6.6.

2.2.4. Discussion Pages: A *discussion* tab at the top of each article takes readers to its *Talk* page, representing a forum for discussions (often longer than the article itself) as to how it might be criticized, improved or extended in the future. For example, the talk page of the *Library* article, *Talk:Library*, contains the following observations, among many others:

> **location?**
>
> Libraries can also be found in churches, prisons, hotels etc. Should there be any mention of this? *– Daniel C. Boyer 20:38, 10 Nov 2003*
>
> Libraries can be found in many places, and articles should be written and linked. A wiki article on libraries can never be more of a summary, and will always be expandable – *DGG 04:18, 11 September 2006*

There are talk pages for other aspects of Wikipedia's structure, such as templates and categories, as well as user talk pages that editors use to communicate with each other. These pages are a unique and interesting feature of Wikipedia not replicated in traditional encyclopedias. They have been mined for determining quality metrics of Wikipedia edits [Emigh *et al.* 2005; Viégas *et al.* 2007] but have not been yet employed for any tasks discussed in this paper—perhaps because of their unstructured nature.

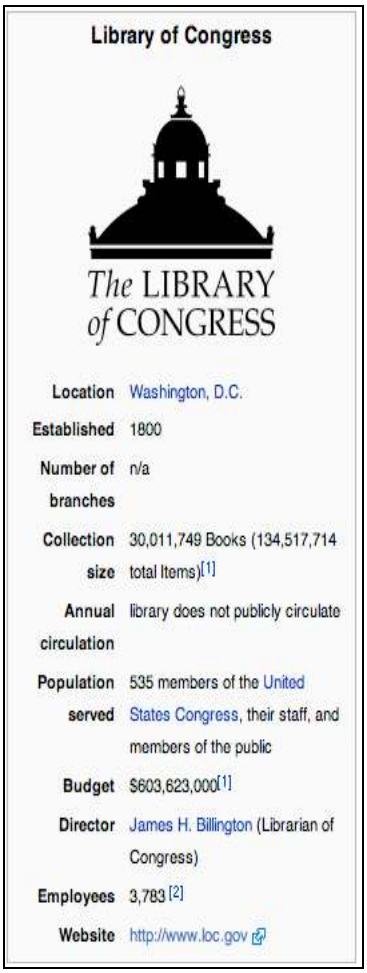

Figure 2. Infobox for the Library of Congress

2.2.5 Edit histories: To the right of the *discussion* tab is a *history* tab that takes readers to each article's editing history. This contains the name or pseudonym of every editor, with the changes they made. From the revision history of *Library* we can see that this article was created on 9 November 2001 in the form of a short note—which, in fact, bears little relationship to the current version—and has been edited about 1500 times since. Recent edits add new links and new entries to lists; indicate possible vandalism and its reversal; correct spelling mistakes; and so on.

Articles and related pages	5,460,000
redirects	2,970,000
disambiguation pages	110,000
Lists and stubs	620,000
bona-fide articles	1,760,000

Categories	390,000
Templates	174,000
infoboxes	9,000
other	165,000

Links	
between articles	62,000,000
between category and subcategory	740,000
between category and article	7,270,000

Table 1. Content of English Wikipedia.

Analyzing editing history is an interesting research area its own right. For example, Viégas [2004] describes how history pages can be mined to discover collaboration patterns. Nelken and Yamangil [2008] discuss several ways of utilizing the unique properties of history pages as a corpus for extracting lexical errors called *eggcorns*, e.g. <*rectify*, *ratify*>, as well as phrases that can be dropped to compress sentences, a useful component of automatic text summarization.

It is natural to ask whether the content of individual articles converges in some semantic sense, staying stable despite continuing edits. Thomas and Amit [2007] call the information in a Wikipedia article "justified" if, after going through the community process of discussion, repeated editing, and so on, it has reached a stable state. They found that articles do, in general, become stable, but that it is difficult to predict where in its journey towards maturity a given article is at any point in time. They also point out that although information about an article's edit history might indicate its likely quality, mining systems invariably ignore it.

Table 1 breaks down the number of different pages and connections in the English version at the time of writing. There are almost 5.5 million pages in the section dedicated to articles. Most are redirects. Many others are disambiguation pages, lists (which group related articles but do not provide explanatory text themselves) and stubs (incomplete articles with fewer than 30 words or at least one incoming link from elsewhere in Wikipedia). Removing all these leaves about 1.8 million bona-fide articles, each with an edit history and most with some content on their discussion page. The articles are organized into 400,000 different categories and augmented with 170,000 different templates. They are densely interlinked, with 62 million connections—an average of 25 incoming and 25 outgoing links from each article.

2.3. Perspectives on Wikipedia

Wikipedia is a rich resource with several different broad functionalities. We will see in subsequent sections that researchers have developed sophisticated mining techniques with which they can identify, isolate and utilize these different perspectives. Here we introduce the most important examples.

2.3.1 Wikipedia as an encyclopedia: The first and most obvious usage for Wikipedia is exactly what it was intended as: an encyclopedia. Ironically, this is the very application that has generated most doubt and cynicism. As noted above, the open editing policy has led many to doubt its authority. Denning *et al.* [2005] provide a good review of early concerns. They conclude that, while Wikipedia is an interesting example of large-scale collaboration, its use as an information source is risky. Their core argument is the lack of formal expert review procedures, which gives rise to two key issues: accuracy within articles, and bias of coverage across them.

Accuracy within articles is investigated by Giles [2005], who compares randomly selected scientific Wikipedia articles with their equivalent entries in Encyclopedia Britannica. Both sources were equally prone to significant errors, such as misinterpretation of important concepts. More subtle errors, however, such as omissions or misleading statements, were more common in Wikipedia. In the 41 articles reviewed there were 162 mistakes in Wikipedia versus 123 for Britannica. Britannica Inc. attacked Giles' study as "fatally flawed"[8] and demanded a retraction; Nature defended itself and declined to retract.[9] Ironically, while Britannica's part in the debate has been polemical and plainly biased, Wikipedia provides objective coverage on the controversy in its article on *Encyclopedia Britannica*.

Several authors have developed metrics that evaluate the quality of Wikipedia articles based on such features as number of authors, number of edits, internal and external linking, and article size, e.g. Lih [2004] and Wilkinson and Huberman [2007]; article stability, e.g. Dondio *et al.* [2006]; and the amount of conflict an article generates, e.g. Kittur [2007]. Emigh and Herring [2005] perform a genre analysis on Wikipedia using corpus linguistic methods to determine "features of formality and informality," and claim that its degree of post-production editorial control produces entries as standardized as those in traditional print encyclopedias. Viégas *et al.* [2007] claim that overall coordination and organization, one of the fastest growing areas of Wikipedia, ensures great resilience to malicious editing despite high traffic; they highlight in particular the role played by discussion pages.

---

[8] *http://www.corporate.britannica.com/britannica_nature_response.pdf*
[9] *http://www.nature.com/press_releases/Britannica_response.pdf*

So much for accuracy. A second issue is *bias of coverage*. Wikipedia is edited by volunteers, who naturally apply more effort to describing topics that pique their interest. For example, there are 600 different articles dedicated to the *The Simpsons* cartoon. In contrast, there are half as many pages about the namesake of the cartoon's main character, the Greek poet *Homer*, and all the literary works he created and inspired. Lih [2004] shows that Wikipedia's content, and therefore bias, is also driven to a large extent by the press. Milne *et al.* [2006] identify a bias towards concepts that are general or introductory, and therefore more relevant to "everyman."

2.3.2. Wikipedia as corpus: Large text collections are useful for creating language models that capture particular characteristics of language use. For example, the language in which a text is written can be determined by analyzing the statistical distribution of the letter n-grams it contains [Cavnar and Trenkle 1994], whereas word co-occurrence statistics are helpful in tasks like spelling correction [Mays *et al.* 1991]. Aligned text corpora in different languages are extremely useful in machine translation [Brown *et al.* 1993]. Extensive coverage and high quality of the corpus is a crucial criterion in the success of such applications. While the web has enabled rapid acquisition of large text corpora, their quality leaves much to be desired, due to spamming and the varying format of websites. In particular, manually annotated corpora and aligned multilingual corpora are still rare and in high demand.

Wikipedia provides a plethora of well-written and well-formulated articles—several gigabytes in the English version alone—that can easily be separated from other parts of the website. The Simple Wikipedia is significantly smaller, but its articles are written for non-English speakers and do not contain complex sentences. This makes automatic linguistic processing easier, and some researchers focus on Simple Wikipedia for their experiments [Ruiz-Casado *et al.* 2005; Toral and Muños 2006]. Many researchers take advantage of the large number of definitions in Wikipedia for question answering (Section 4.3) and automatic extraction of semantic relations (Section 5.1). Section 2.2.5 mentions how Wikipedia history pages can be used as a corpus for training text summarization algorithms, as well as for determining the quality of the articles themselves.

Wikipedia also contains annotations in the form of targeted hyperlinks. Consider the following two sentences from the article about the Formula One team named McLaren.

1. The [[Kiwi (people)|Kiwi]] made the team's Grand Prix debut at the 1966 Monaco race.
2. Original McLaren [[Kiwi|kiwi]] logo; a New Zealand icon.

In the first case the word *kiwi* links to *Kiwi* (*people*); in the second, to *Kiwi*, the article describing the bird. This mark-up is nothing more or less than word sense annotation.

Mihalcea [2007] shows that Wikipedia is a full fledged alternative to manually sense-tagged corpora. Section 3.2 discusses research that makes use of these annotations for word sense disambiguation and computing the semantic similarity between words.

Although the exploration of Wikipedia as a source of multilingual aligned corpora has only just begun, its links between description of concepts in different languages have been exploited for cross-language question answering [Ferrández *et al.* 2007] and automatic generation of bilingual dictionaries [Erdmann *et al.* 2008]. This is further discussed in Section 3.4, while Section 4.3 investigates Wikipedia's potential for multilingual information retrieval.

2.3.3 Wikipedia as a thesaurus: There are many similarities between the structure of traditional thesauri and the ways in which Wikipedia organizes its content. As noted, each article describes a single concept, and its title is a succinct, well-formed phrase that resembles a term in a conventional thesaurus. If article names correspond to manually defined terms, links between them correspond to relations between terms, the building blocks of thesauri. The international standard for thesauri (ISO 2788) specifies four kinds of relation:

- Equivalence: USE, with inverse form USE FOR
- Hierarchical: broader term (BT), with inverse form narrower term (NT)
- Any other kind of semantic relation (RT, for related term).

Wikipedia redirects provide precisely the information expressed in the equivalence relation. As noted, they are a powerful way of dealing with word variations such as abbreviations, equivalent expressions and synonyms. The hierarchical relations (broader and narrower terms) are reflected in Wikipedia's category structure. Hyperlinks between articles capture other kinds of semantic relation. (Restricting consideration to mutual cross-links eliminates many of the more tenuous associations.)

As we will see, researchers compare Wikipedia with manually created domain-specific thesauri and augment them with knowledge from it (Section 3.2.3). Redirects turn out to be very accurate and can safely be added to existing thesauri without further checking. Wikipedia also has the potential to contribute new topics and concepts, and can be used as a source of suggestions for thesaurus maintenance. Manual creation of scope notes is a labor-intensive aspect of traditional thesauri. Instead, the first paragraph of a Wikipedia article can be extracted as a description of the topic, backed up by the full article should more explanation be required. Finally, Wikipedia's multilingual nature allows thesauri to be translated into other languages.

2.3.4. Wikipedia as a database: Wikipedia contains a massive amount of highly structured information. Several projects (notably DBpedia, discussed in Sections 5.2 and

6.6) extract this and store it in formats accessible to database applications. The aim is two-fold: to allow users to pose database-style queries against datasets derived from Wikipedia, and to facilitate linkage with other datasets on the web. Some projects even aim to extract database-style facts directly from the text of Wikipedia articles, rather than from infoboxes. Furthermore, disambiguation and redirect pages can be turned into a relational database that contains tables for *terms*, *concepts*, *term concept relationships* and *concept relationships* [Gregorowicz and Kramer 2006].

Another idea is to bootstrap fact extraction from articles by using the content of infoboxes as training data and applying machine learning techniques to extract even more infobox-style information from the text of other articles. This allows infoboxes to be generated for articles that do not yet have them [Wu and Weld 2007]. Related techniques can be used to clean up the underlying infobox data structure, with its proliferation of individual templates.

2.3.5 Wikipedia as an ontology: Articles can be viewed as ontology elements, for which the URIs of Wikipedia entries serve as surprisingly reliable identifiers [Hepp et al. 2006]. Of course, true ontologies also require concept nodes to be connected by informative relations, and in Section 6 we will see researchers mine such relations in a host of innovative ways from Wikipedia's structure—including redirects, hyperlinks (both incoming and outgoing, as well as the anchor text), category links, category names and infoboxes, and even raw text, as well as experimenting with adding relations to and from other resources such as WordNet and Cyc.

From this viewpoint Wikipedia is arguably by far the largest living ontological structure available today, with its distinctive Wiki technology serving as a large-scale collaborative ontology development environment. Some researchers are beginning to mix traditional mining techniques with possibly more far-sighted attempts to encourage Wikipedia editors themselves in directions that might bear ontological fruit.

2.3.6 Wikipedia as a network structure: Wikipedia can be viewed as a hyperlinked structure of web pages, a microcosm of the web. Standard methods of analyzing the network structure can then be applied [Bellomi and Bonato 2005]. The two most prominent techniques used for web analysis are PageRank, which underpins Google's success [Brin and Page 1998], and the HITS algorithm [Kleinberg 1998]. Bellomi and Bonato [2005] applied both of these to Wikipedia and discerned some interesting underlying cultural biases (as of April 2005). These authors conclude that PageRank and HITS seem to identify different kinds of information. They report that according to the HITS authority metric, space (in the form of political geography) and time (in the form of both time spans and landmark events) are the primary organizing categories for Wikipedia

articles. Within these, information tends to be organized around famous people, common words, animals, ethnic groups, political and social institutions, and abstract concepts such as music, philosophy, and religion.

In contrast, the most important articles according to PageRank include an overwhelming number of concepts tightly related to religion. For example, *Pope*, *God* and *Priest* were the highest-ranking nouns, as compared to *Television*, *Scientific classification*, and *Animal* for HITS. They found that PageRank seemed to transcend recent political events to give a wider historical and cultural perspective in weighting geographic entities. It also tends to bring out a global rather than a Western perspective, both for countries and cities and for historical events. HITS reveals a strong bias towards recent political leaders, whereas people with high PageRank scores tend to be ones with an impact on religion, philosophy and society. It would be interesting to see how these trends have evolved in the three years since the publication of this work.

An alternative to PageRank and HITS is the Green method [Duffy 2001], which Ollivier and Senellart [2007] applied to Wikipedia's hyperlink network structure in order to find related articles. This method, which is based on Markov Chain theory, is related to the topic-sensitive version of PageRank introduced by Haveliwala [2003]. Given a target article, one way of finding related articles is to look at nodes with high PageRank in its immediate neighborhood. For this a topic-sensitive measure like Green's is more appropriate than the global PageRank.

The Wikipedia category graph also forms a network structure. Zesch and Gurevych [2007] showed that it is a scale-free, small-world graph, like other semantic networks such as WordNet. They adapted WordNet-based measures of semantic relatedness to use the Wikipedia category graph instead, and found that they work well—at least for nouns. They suggest that this, coupled with Wikipedia's multilingual nature, may enable natural language processing algorithms to be transferred to languages that lack well-developed semantic WordNets.

2.4. Obtaining Wikipedia data

Wikipedia is based on the MediaWiki software. As an open source project, its entire content is easily obtainable. It is available in the form of large XML files and database dumps that are released sporadically, from several days to several weeks apart.[10] The full content (without revision history or images) of the English version of Wikipedia occupies 18 Gb of uncompressed data at the time of writing. There are several tools for extracting information from these files, which are discussed in Section 7.

[10] *http://download.wikimedia.org/wikipedia*

Instead of obtaining the database directly, specialized web crawlers have been developed to download the entire content of Wikipedia. Bellomi and Bonato [2005] scanned the *All pages* index section, which contains a complete list of the pages exposed on the website. Pages that do not contain a regular article were identified by testing for specific patterns in the URL, and discarded. Wikipedia's administrators prefer the use of the database dumps, however, to minimize the strain placed on their services.

3 SOLVING NATURAL LANGUAGE PROCESSING TASKS

Natural language processing applications fall into two major groups: i) those relying on symbolic methods, where the system utilizes a manually encoded repository of human language, and ii) statistical methods, which infer properties of language by processing large text corpora. The problem with the former is a dearth of high-quality knowledge bases. Even the lexical database WordNet, which, as the largest of its kind, receives substantial attention [Fellbaum 1998], has been criticized for low coverage—particularly of proper names—and high sense proliferation [Mihalcea and Moldovan 2001; Ponzetto and Strube 2007a]. Initial enthusiasm with statistical methods somewhat faded once they hit an upper performance bound that is hard to improve upon unless they are combined with symbolic elements [Klavans and Resnik 1996]. Several research groups simultaneously discovered Wikipedia as an alternative to WordNet. Direct comparison of their performance on the same task has shown that Wikipedia can be employed in a similar way and significantly outperforms WordNet on various tasks [Strube and Ponzetto 2006]. This section describes research in the four language processing tasks to which Wikipedia has been successfully applied: semantic relatedness (Section 3.1), word sense disambiguation (Section 3.2), co-reference resolution (Section 3.3) and multilingual alignment (Section 3.4).

3.1 Semantic relatedness

Semantic relatedness quantifies the similarity between two concepts, e.g. *doctor* and *hospital*. Budanitsky and Hirst [2001] differentiate between semantic *similarity*, where only predefined taxonomic relations are used to compute similarity, and semantic *relatedness*, where other relations like *has-part*, *is-made-of* are used as well. Semantic relatedness can be also quantified by statistical methods without requiring a manually encoded taxonomy, for example by analyzing term co-occurrence in a large corpus [Resnik 1995; Jiang and Conrath 1997].

To evaluate automatic methods for estimating semantic relatedness, the correlation coefficient between machine-assigned scores and those assigned by human judges is computed. Three standard datasets are available for evaluation:

- Miller and Charles' [1991] list of 30 noun pairs, which we denote by M&C;
- Rubenstein and Goodenough's [1965] 65 synonymous word pairs, R&G,
- [Finkelstein *et al.* 2002]'s collection of 353 word pairs (WordSimilarity-353), WS-353.

The best pre-Wikipedia result for the first set was a correlation of 0.86, achieved by Jiang and Conrath [1997] using a combination of statistical measures and taxonomic analysis derived from WordNet. For the third, Finkelstein *et al.* [2002] achieved 0.56 correlation using Latent Semantic Analysis. The discovery of Wikipedia began a new era of competition.

Strube and Ponzetto [2006] and Ponzetto and Strube [2007a] re-calculated several measures developed for WordNet using Wikipedia's category structure. The best performing metric on most datasets was Leacock and Chodorow's [1998] normalized path measure:

$$lch(c_1,c_2) = -\log \frac{length(c_1,c_2)}{2D},$$

where *length* is the number of nodes on the shortest path between nodes c_1 and c_2, and D is the maximum depth of the taxonomy. WordNet-based measures outperform Wikipedia-based ones on the small datasets M&C and R&G, but on WS-353 Wikipedia wins by a large margin. Combining similarity evidences from Wikipedia and WordNet using a SVM to learn relatedness from the training data yielded the highest correlation score of 0.62 on a designated "testing" subset of WS-353.

Strube and Ponzetto remark that WordNet's sense proliferation was responsible for its poor performance on WS-353. For example, when computing the relatedness of *jaguar* and *stock*, the latter is interpreted in the sense of animals kept for use or profit rather than in the sense of *market*, which people find more intuitive. WordNet's fine sense granularity has been also criticized in word sense disambiguation (Section 3.2.1). The overall conclusion is that Wikipedia can serve AI applications in the same way as hand-crafted knowledge resources.

Zesch *et al.* [2007] perform similar experiments with the German Wikipedia, which they compare to GermaNet on three datasets including the translated M&C. The performance of Wikipedia-based measures was inconsistent, and, like Strube and Ponzetto [2006], they obtained best results by combining evidence from GermaNet and Wikipedia.

Ponzetto and Strube [2007a] investigate whether performance on Wikipedia-based relatedness measures changes as Wikipedia grows. After comparing February 2006, September 2006 and May 2007 versions they conclude that the relatedness measure is robust. There was no improvement, probably because new articles were unrelated to all

words in the evaluation datasets. A Java API is available for those wishing to experiment with these techniques [Ponzetto and Strube [2007c].[11]

Gabrilovich and Markovitch [2007] develop Explicit Semantic Analysis (ESA) as an alternative to the well-known Latent Semantic Analysis. They use a centroid-based classifier to map input text to a vector of weighted Wikipedia articles. For example, for *Bank of Amazon* the vector contains *Amazon River*, *Amazon Basin*, *Amazon Rainforest*, *Amazon.com*, *Rainforest*, *Atlantic Ocean Brazil*, etc. To compute semantic relatedness between two terms, they compute the cosine similarity of their vectors. This significantly outperforms Latent Semantic Analysis on WS-353, with an average correlation of 0.75. With the same technique, the Open Directory Project[12] achieves a 0.65 correlation, indicating that Wikipedia's quality is greater. The mapping developed in this work has been successfully utilized for text categorization (Section 4.4).

While Gabrilovich and Markovitch [2007] use the full text of Wikipedia articles to establish relatedness between terms, Milne [2007] analyses just the internal hyperlinks that appear, arguing that Wikipedia's link structure bears much significant information about concepts. To compute the relatedness between two terms they are first mapped to corresponding Wikipedia articles, and then vectors are created containing the links to other Wikipedia articles that occur in these articles. For example, a sentence like *Bank of America is the largest commercial <bank> in the <United States> by both <deposits> and <market capitalization>* contributes four links to the vector. Each link is weighted by the inverse number of times it is linked from other Wikipedia articles—the less common the link, the higher its weight. For example, *market capitalization* receives higher weight than *United States* and thus contributes more to the semantic relatedness.

Disambiguation is a serious challenge for this technique. Strube and Ponzetto [2006] choose the most likely meaning from the order in which entries occur in Wikipedia's disambiguation pages; Gabrilovich and Markovitch [2007] avoid disambiguation entirely by simultaneously associating a term with several Wikipedia articles. However, Milne's [2007] approach hinges upon correct mapping of terms to Wikipedia articles. When terms are manually disambiguated, a correlation of 0.72 is achieved for WS-353. Automatic disambiguation that simply selects whatever meaning produces the greatest similarity score is only 0.45, showing that unlikely senses often produce greater similarity than common ones.

Milne and Witten [2008a] disambiguate term mappings automatically using three features. One is the conditional probability of the sense given the term, according to the Wikipedia corpus (discussed further in Section 3.2.1). For example, the term *leopard*

[11] *http://www.eml-r.org/english/research/nlp/download/jwordnetsimilarity.php*

most often links to the animal description rather than the eponymous Mac operating system. They also analyze how commonly two terms appear in Wikipedia as a collocation. Finally, they replace the vector-based similarity metric described above by a measure inspired by Cilibrasi and Vitanyi's [2002] Normalized Google Distance, which is based on term occurrences in web pages, but using Wikipedia's links rather than Google's search results. The semantic similarity of two terms is determined by the sum of these three values—conditional probability, collocation and similarity.

This technique achieves 0.69 correlation with human judgments on WS-353, not far off Gabrilovich and Markovitch's [2007] figure for ESA. However, it is far less computationally intensive because only links are analyzed, not the entire Wikipedia text. Further analysis of the results shows that performance is even higher on terms that are well defined in Wikipedia.

Table 2 summarizes the results of the similarity metrics that we have described, using the same datasets and evaluation technique. ESA is best, with WLM not far behind and WikiRelate the lowest. The astonishingly high correlation with human performance that these techniques obtain was well out of reach in pre-Wikipedia days. This is an important advance, because—as we will see when discussing information retrieval and extraction—automatic computation of semantic similarity helps with many natural language processing tasks.

3.2 Word sense disambiguation

Techniques for word sense disambiguation—i.e., resolving polysemy—use a dictionary or thesaurus that defines the inventory of possible senses [Ide and Veronis 1998]. Wikipedia provides an alternative resource. Each article describes a concept that is a possible sense for words and phrases that denote it, whether by redirection, via a disambiguation page, or as anchor text that links to the article.

The terms to be disambiguated may either appear in plain text or in an existing knowledge base (thesaurus or ontology). The former situation is more complex because the context is less clearly defined. Consider the example in Figure 3. Even human readers cannot be sure of the intended meaning of *wood* from the sentence alone, but a diagram showing semantically related words in WordNet puts it into context and makes it clear that the meaning is *the trees and other plants in a large densely wooded area*, rather than *the hard fibrous lignified substance under the bark of trees*. This highlights the main idea behind disambiguation: identify the context and analyze which of the possible senses fits it best.

---

[12] *http://www.dmoz.org*

Method	M&C	R&G	WS-353
WordNet [Strube and Ponzetto, 2006]	0.82	0.86	full: 0.36 test: 0.38
WikiRelate! [Ponzetto and Strube, 2007]	0.49	0.55	full: 0.49 test: 0.62
ESA [Gabrilovich and Markovitch, 2007]	0.73	0.82	0.75
WLVM [Milne, 2007]	n/a	n/a	man: 0.72 auto: 0.45
WLM [Milne and Witten, 2008]	0.70	0.64	0.69

Table 2. Overview of semantic relatedness methods.

We first cover techniques for disambiguating phrases in text to Wikipedia articles, then examine the important special case of named entities, and finally show how disambiguation is used to map manually created knowledge structures to Wikipedia.

3.2.1. Disambiguating phrases in running text: Discovering the intended senses of words and phrases is an essential stage in every natural language application, otherwise full "understanding" cannot be claimed. WordNet is a popular resource for word sense disambiguation, but success has been mixed [Voorhees 1998]. One reason is that the task is demanding because "linguistic [disambiguation] techniques must be essentially perfect to help" [Vorhees 1998]; another is that WordNet defines word senses with such fine granularity that even human annotators struggle to differentiate them [Edmonds and Kilgariff 1998]. The two are related, because fine sense granularity makes disambiguation more difficult. In contrast, Wikipedia defines only those senses on which its contributors reach consensus, and include an extensive description of each one rather than WordNet's brief gloss. Substantial advances have been made since it was discovered as a resource for disambiguation.

Mihalcea [2007] use Wikipedia articles as a source of sense-tagged text to form a training corpus for supervised disambiguation. They follow the evaluation methodology developed by SIGLEX, the Association for Computational Linguistics' Special Interest Group on the Lexicon.[13] For each example they collect its occurrences as link anchors in Wikipedia. For example, the term *bar* is linked to *bar* (*establishment*) and *bar* (*music*), each of which corresponds to a WordNet synset—that is, a set of synonymous terms representing a particular meaning of *bar*. The results show that a machine learning approach trained on Wikipedia sentences in which both meanings of *bar* occur clearly outperforms two simple baselines.

He could see wood around the house.

Figure 3. What is the meaning of *wood* in both examples?

This work uses Wikipedia solely as a resource to disambiguate words or phrases into WordNet synsets. Mihalcea and Csomai [2007] go further, using Wikipedia's content as a sense inventory in its own right. They disambiguate terms—words or phrases—that appear in plain text to Wikipedia articles, concentrating exclusively on "important" concepts. They call this process *wikification* because it simulates how Wikipedia authors manually insert hyperlinks when writing articles. There are two stages: extraction and disambiguation. In the first, terms that are judged important enough to be highlighted as links are identified in the text. Only terms occurring at least five times in Wikipedia are considered, and *likelihood* of a term being a hyperlink is estimated by expressing the number of articles in which a given word or phrase appears as anchor text as a proportion of the total number of articles in which it appears. All terms whose likelihood exceeds a predefined threshold are chosen, which yields an F-measure of 55% on a subset of manually annotated Wikipedia articles.

In the second stage these terms are disambiguated to Wikipedia articles that capture the intended sense. For example, in the sentence *Jenga is a popular beer in the bars of Thailand* the term *bar* corresponds to the *bar* (*establishment*) article. Given a term, those articles for which it is used as anchor text in the Wikipedia are candidate senses. Best results are achieved by a machine learning approach in which Wikipedia's already-annotated articles serve as training data. Features—like part-of-speech tag, local context of three words to the left and right, and their part-of-speech tags—are computed for each ambiguous term that appears as anchor text of a hyperlink. A Naïve Bayes classifier is then applied to disambiguate unseen terms. Csomai and Mihalcea [2007] report an F-measure of 87.7% on 6,500 examples, and go on to demonstrate that linking educational material to Wikipedia articles in this manner improves the quality of knowledge that people acquire when reading the material, and decreases the time taken.

---

[13] *http://www.senseval.org*

In a parallel development, Wang *et al.* [2007] use a fixed-length window to identify terms in a document that match the titles of Wikipedia articles, eliminating matches subsumed by longer ones. They disambiguate the matches using two methods. One works on a document basis, seeking those articles that are most similar to the original document according to the standard cosine metric between TF×IDF-weighted word frequency vectors. The second works on a sentence basis, computing the shortest distance between the candidate articles for a given ambiguous term and articles corresponding to any non-ambiguous terms that appear in the same sentence. The distance metric is 1 if the two articles link to each other; otherwise it is the number of nodes along the shortest path between two Wikipedia categories to which they belong, normalized by the maximum depth of the category taxonomy. The result is the average of the two techniques (if no unambiguous articles are available, the similarity technique is applied by itself). Wang *et al.* do not compare this method to other disambiguation techniques directly. They do, however, report the performance of text categorization before and after synonyms and hyponyms of matching Wikipedia articles, and their related terms, were added to the documents. The findings were mixed, and somewhat negative.

Medelyan *et al.* [2008] use Mihalcea and Csomai's [2007] wikification strategy with a different disambiguation technique. Document terms with just one match are unambiguous, and their corresponding articles are collected and used as "context articles" to disambiguate the remaining terms. This is done by determining the average semantic similarity of each candidate article to all context articles identified for the document. The semantic similarity of a pair of articles is obtained from their incoming links as described by Milne and Witten [2008a] (see Section 3.1). Account is also taken of the conditional probability of a sense given the term, according to the Wikipedia corpus (proposed by Mihalcea and Csomai [2007] for a baseline). For example, the term *jaguar* links to the article *Jaguar cars* in 466 out of 927 cases, thus its conditional probability is 0.5. The resulting mapping is the one with the largest product of semantic similarity and conditional probability. This achieves an F-measure of 93% on 17,500 mappings in manually annotated Wikipedia articles.

Milne and Witten [2008b] extend this approach using machine learning. Rather than extracting terms and then disambiguating them, they allow a term's possible mappings to influence whether it should be adjudged an important concept for the document. Conditional probability of a mapping, its semantic similarity to other context articles, and other features are combined in a machine learning classifier, bagged decision trees, which determines a probability figure for each mapping. More than one Wikipedia article

can be chosen for a given document term, which improves recall at the expense of a slight decrease in precision, raising the F-measure from 93% to 97% on the same data.

3.2.2. Disambiguating named entities: Phrases referring to named entities, which are proper nouns such as geographical and personal names, and titles of books, songs and movies contribute to the largest part of our vocabulary. Wikipedia is recognized as the largest available resource of such entities. It has become a platform for discussing current news, and contributors put issues into encyclopedic context by relating them to historical events, geographic locations and significant personages, thereby increasing the coverage of named entities. Here we describe three approaches that focus specifically on linking named entities appearing in text or in search queries to corresponding Wikipedia articles. Techniques for recognizing named entities in Wikipedia itself are summarized in Section 5.3.

Bunescu and Paşca [2006] disambiguate named entities in search queries in order to group search results by the corresponding senses. They first create a dictionary of 500,000 entities that appear in Wikipedia, and add redirects and disambiguated names to each one. If a query contains a term that corresponds to two or more entries, they choose the one whose Wikipedia article has the greatest cosine similarity with the query. If the similarity scores are too low they use the category to which the article belongs instead of the article itself. If even this falls below a predefined threshold they assume that no mapping is available. The reported accuracies are between 55% and 85% for members of Wikipedia's *People by occupation* category, depending on the model and experimental data employed.

Cucerzan [2007] identifies and disambiguates named entities in text. Like Bunescu and Paşca [2006], he first extracts a vocabulary from Wikipedia. It is divided into two parts, the first containing surface forms and the second the associated entities, along with contextual information about them. The surface forms are titles of articles, redirects, and disambiguation pages, and anchor text used in links. This yields 1.4 million entities, with an average of 2.4 surface forms each. Further <named entity, tag> pairs are extracted from Wikipedia list pages—e.g., *Texas* (*band*) receives a tag *LIST_band name etymologies*, because it appears in the list with this title—yielding a further 540,000 entries. Categories assigned to Wikipedia articles describing named entities serve as tags too, yielding 2.65 million entries. Finally a context for each named entity is collected—e.g., parenthetical expressions in its title, phrases that appear as link anchors in the article's first paragraph of the article, etc.—yielding 38 million <named entity, context> pairs.

To identify named entities in text, capitalization rules indicate which phrases are surface forms of named entities. Co-occurrence statistics generated from the web by a search engine help to identify boundaries between them (e.g. *Whitney Museum of American Art* is a single entity, whereas *Whitney Museum in New York* contains two). Lexical analysis is used to collate identical entities (e.g., *Mr. Brown* and *Brown*), and entities are tagged with their type (e.g., *location*, *person*) based on statistics collected from manually annotated data. Disambiguation is performed by comparing the similarity of the document in which the surface form appears with Wikipedia articles that represent all named entities that have been identified in it, and their context terms, and choosing the best match. Cucerzan [2007] achieves 88% accuracy on 5,000 entities appearing in Wikipedia articles, and 91% on 750 entities appearing in news stories.

Kazama and Torisawa [2007] recognize and classify entities but do not disambiguate them. Their work resembles the methods described above. Given a sentence, their goal is to extract all n-grams representing Wikipedia articles that correspond to a named entity and assign a type to it. For example, in the sentence *Rare Jimmy Hendrix song draft sells for almost $17,000* they identify *Jimmy Hendrix* as an entity of type *musician*. To determine the type they extract the first noun phrase following the verb *to be* from the Wikipedia article's first sentence, excluding phrases like *kind of*, *type of*—e.g., *guitarist* in *Jimmy Hendrix was a guitarist*. Recognition is a supervised tagging process based on standard features such as surface form and part of speech tag, augmented with category labels extracted from Wikipedia and a gazetteer. An F-measure of 88% was achieved on a standard set of 1000 training and 220 development and testing documents.

Cucerzan [2007] and Kazama and Torisawa [2007] report similar performance, while Bunescu and Paşca's [2006] results seem slightly worse. However, comparison is unreliable because different datasets are used. Accuracy also depends on the type of the named entity.

3.2.3. Disambiguating thesaurus and ontology terms: Wikipedia's category and link structure contains the same kind of information as a domain-specific thesaurus, as illustrated by Figure 4, which compares it to the agricultural thesaurus Agrovoc [1995]. Whereas in Section 3.1.2 Wikipedia is used as an independent knowledge base, it can also be used to extend and improve existing resources. For example, if it were known that *cardiovascular system* and *circulatory system* in Figure 4 refer to the same concept, the synonym *blood circulation* could be added to Agrovoc. The major problem is to establish a mapping between Wikipedia and other resources, disambiguating situations that support multiple mappings.

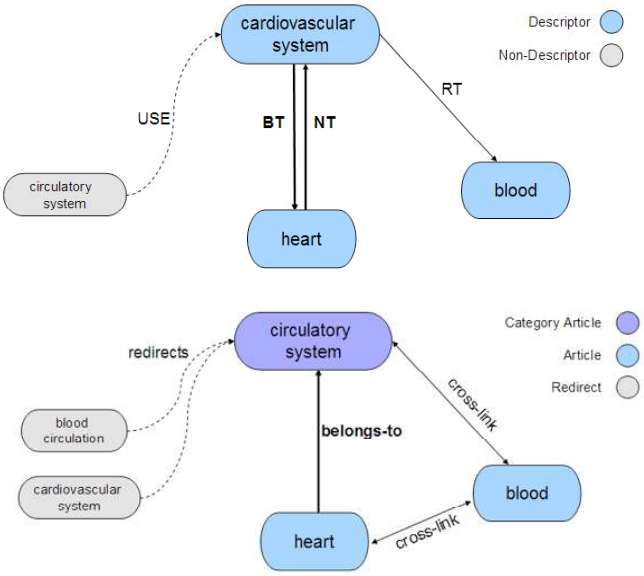

Figure 4. Comparison of organization structure in Agrovoc and Wikipedia.

Ruiz-Casado *et al.* [2005] map Wikipedia articles to WordNet. They work with the Simple Wikipedia,[14] a reduced version that contains easier words and shorter sentences, intended for people learning English. WordNet synsets cluster word senses so that homonyms can be identified. If a Wikipedia article matches several WordNet synsets, the appropriate one is chosen by computing the similarity between the Wikipedia entry word-bag and the WordNet synset gloss. This technique achieves 84% accuracy, when dot product similarity of stemmed word vectors is applied. The problem is that as Wikipedia grows, so does ambiguity. For instance even the Simple Wikipedia contains the article *Cats* (*musical*), which is absent from WordNet. The mapping technique must be able to deal with absent items as well as polysemy in both resources.

Overell and Rüger [2006] disambiguate place names mentioned in Wikipedia to locations in gazetteers. Instead of semantic similarity they develop geographically-based disambiguation methods. One seeks a minimum bounding box enclosing the location being disambiguated and other place names that are mentioned in the same context, using geographical coordinates from the gazetteer. Another analyzes the place name's referent; for example, if the surface form *Ontario* is mapped to *Ontario, Canada*, then *London, Ontario* can be mapped to *London, Canada*. Best results were achieved by combining the minimum bounding box method with "importance," measured by population size.

[14] http://simple.wikipedia.org

An F-measure of 80% was achieved on a test set with 1,700 locations and 12,275 non-locations.

Overell and Rüger [2007] extend this approach by creating a co-occurrence model for each place name. They map place names to Wikipedia articles, collect their redirects as synonyms, and gather the anchor text of links to these articles. This yields different ways of referring to the same place, e.g., {*Londinium* → *London*} and {*London*, *UK* → *London*}. Next they collect evidence from Wikipedia articles: geographical coordinates, and location names in subordinate categories. They also mine Placeopedia, a mash-up website that connects Wikipedia with Google Maps. Together, these techniques recognize 75% of place names and map them to geographical locations with an accuracy of between 78 and 90%.

Milne *et al*. [2007] investigate whether domain-specific thesauri can be obtained from Wikipedia for use in natural language applications within restricted domains, comparing it with Agrovoc, a manually built agricultural thesaurus. On the positive side, Wikipedia article titles cover the majority of Agrovoc terms that were chosen by professional indexers as index terms for an agricultural corpus, and its redirects correspond closely with Agrovoc's synonymy relation. However, neither category relations nor (mutual) hyperlinks between articles correspond well with Agrovoc's taxonomic relations. Instead of extracting new domain-specific thesauri from Wikipedia they examine how existing ones can be improved, using Agrovoc as a case study [Medelyan and Milne 2008]. Given an Agrovoc descriptor, they collect semantically related terms from the Agrovoc hierarchy as context terms and map each one to the Wikipedia articles whose conditional probability (as explained in Section 3.2.1) is greatest. Then they compute the semantic similarity of each candidate mapping to this set of context articles. Manual evaluation of a subset with 400 mappings shows an average accuracy of 92%. The results are slightly better if there are fewer than four possible mappings and remain stable at 88% if there are ten or more.

Medelyan and Legg [2008] map terms from the Cyc ontology to Wikipedia articles using the disambiguation approach proposed by Medelyan and Milne [2008]. However, since they draw on the Cyc ontology as part of their disambiguation, and the project can be viewed as a large-scale 'ontology alignment', discussion of it will be postponed to Section 6.5.

There is still far less research on word sense disambiguation using Wikipedia than for WordNet. However, significant advances have been made, and over the last two years the accuracy of mapping documents to relevant Wikipedia articles has improved by one third [Milne and Witten 2008]. Other researchers (such as Wang *et al*. [2007]) use word sense

disambiguation as a part of an application but do not provide an intrinsic evaluation. Furthermore, for fair comparison the same version of Wikipedia and the same training and test set should be used, as has been done for WordNet by SIGLEX (Senseval, cited earlier). Evaluation of named entity extraction is even more complex, with each research group concentrating on different types of entity, e.g. persons or places. Here, extrinsic evaluations may be helpful—e.g., performance on a particular task, for example question answering, before and after integration with Wikipedia. The next section describes an extrinsic evaluation of Wikipedia for co-reference resolution and compares the results with WordNet.

3.3 Co-reference resolution

Natural language understanding tasks such as textual entailment and question answering involve co-reference resolution—identifying which text entities refer to the same concept. Unlike word sense disambiguation, it is not necessary to determine the actual meaning of these entities, but merely identify their connection. Consider the following example from Wikipedia's article on New Zealand:

> **Elizabeth II**, as **the Queen** of New Zealand, is the Head of State and, in **her** absence, is represented by a non-partisan Governor-General. **The Queen** "reigns but does not rule." **She** has no real political influence, and **her** position is essentially symbolic. [emphasis added]

Without knowing that *Elizabeth II* and *the Queen* refer to the same entity, which can be referred to by the pronouns *she* and *her*, the information that can be inferred from this paragraph is limited. To resolve the highlighted co-referent expressions requires linguistic knowledge and world knowledge—that *Elizabeth II* is *the Queen*, and female. Current methods often derive semantic relations from WordNet or mine large corpora using lexical patterns such as *X is a Y* and *Y such as X*. The task can be modeled as a binary classification problem—to determine, for each pair of entities, whether they co-refer or not—and addressed using machine learning techniques, with features such as whether they are semantically related, the distance between them, agreement in number and gender.

The use of Wikipedia for these tasks has been explored in two ways. Ponzetto and Strube [2006a, 2007] analyze its hyperlink structure and text to extract semantic features; whereas Yang and Su [2007] use it as a large semi-structured corpus for mining lexical patterns. They are easy to compare because both use test data from the Message Understanding Conference organized by NIST.

Ponzetto and Strube's [2006, 2007a] main goal is to show that Wikipedia can be used as a fully-fledged lexical and encyclopedic resource, comparable to WordNet but far

		NWIRE			BNEWS		
		R	P	F	R	P	F
Ponzetto and Strube [2006, 2007a]	baseline	56.3	86.7	68.3	50.5	82.0	62.5
	+WordNet	62.4	81.4	70.7	59.1	82.4	68.8
	+Wikipedia	60.7	81.8	69.7	58.3	81.9	68.1
Yang and Su [2007]	baseline	54.5	80.3	64.9	52.7	75.3	62.0
	+sem. related.	57.4	80.8	67.1	54.0	74.7	62.7

Table 3. Performance comparison of two independent techniques on the same datasets.

more extensive. While their work on semantic relatedness (Section 3.1) evaluates Wikipedia intrinsically, co-reference is evaluated extrinsically to demonstrate Wikipedia's utility. As a baseline they re-implement Soon *et al.*'s [2001] method with a set of standard features, such as whether the two entities share the same grammatical feature, or belong to the same WordNet class. Additional features mined from WordNet and Wikipedia are evaluated separately. The WordNet features for two given terms A, e.g. *Elisabeth II*, and B, e.g. *Queen*, are:

- The highest similarity score from all synset pairs to which A and B belong
- The average similarity score.

The Wikipedia analogue to these two features,

- The highest similarity score from all Wikipedia categories to which A and B belong
- The average similarity score,

is augmented by further features:

- Does the first paragraph of the Wikipedia article describing A mention B?
- Does any hyperlink in A's article target B?
- Does the list of categories for A's article contain B?
- What is the overlap between the first paragraphs of the articles for A and B?

The similarity and relatedness scores are computed using various metrics. Feature selection is applied during training to remove irrelevant features for each scenario. The results are included in Table 3, which we will discuss shortly.

Yang and Su [2007] utilize Wikipedia in a different way, assessing semantic relatedness between two entities by analyzing their co-occurrence patterns in Wikipedia. (Pattern matching using the Wikipedia corpus is practiced extensively in information extraction, as described in Section 5). The patterns are evaluated based on positive instances in the training data that serve as seeds. For example, given the pair of co-referents *Bill Clinton* and *president*, and Wikipedia sentences like *Bill Clinton is elected President of the United States* and *The US president, Mr Bill Clinton*; the patterns [*X is elected Y*] and [*Y, Mr X*] are extracted. Sometimes patterns occur in structured parts of Wikipedia like lists and infoboxes—for example, in *United States | Washington, D.C.*, the bar symbol is the pattern. An accuracy measure is used to eliminate patterns that are frequently associated with both negative and positive pairs. Yang and Su [2007] found

that using the 100 most accurate patterns as features did not improve performance over the baseline. However, adding a single feature representing semantic relatedness between the two entities did improve results. Yang and Su use mined patterns to assess relatedness by multiplying together two measures of reliability: the strength of association between each positive seed pair and the pointwise mutual information between the entities occurring with the pattern and by themselves.

Table 3 shows the results that both sets of authors report for co-reference resolution. They use the same baseline, but the implementation was evidently slightly different, for Ponzetto and Strube's yielded a slightly improved F-measure. Ponzetto and Strube's results when features were added from WordNet and Wikipedia are remarkably similar, with no statistical difference between them. These features decrease precision over the baseline on NWIRE by 5 points but increase recall on both datasets, yielding a significant overall gain (1.5 to 2 points on NWIRE and 6 points on BNEWS). Yang and Su improve the F-measure on NWIRE and recall on BNEWS by 2 points. Overall, it seems that Ponzetto and Strube's technique performs slightly better.

These co-reference resolution systems are quite complex, which may explain why no other methods have been described in the literature. We expect further developments in this area.

3.4 Multilingual alignment

In 2006, five years after its inception, Wikipedia contained 100,000 articles for eight different languages. The closest precedent to this unique multilingual resource is the commercial EuroWordNet that unifies seven different languages but covers a far smaller set of concepts—8,000 to 44,000, depending on the language [Vossen *et al.* 1997]. Of course, multilingual vocabularies and aligned corpora benefit any application that involves machine translation.

Adafre and de Rijke [2006] began by generating parallel corpora in order to identify similar sentences—those whose information overlaps significantly—in English and Dutch. First they used a machine translation tool to translate Wikipedia articles and compared the result with the corresponding manually written articles in that language. Next they generated a bilingual lexicon from links between articles on the same topic in different languages, and determined sentence similarity by the number of shared lexicon entries. They evaluated these two techniques manually on 30 randomly chosen Dutch and English Wikipedia articles. Both identified rather a small number of correct sentence alignments: the machine translation had lower accuracy but higher coverage than the lexicon approach. The authors ascribed the poor performance to the small size of the Dutch version but were optimistic about Wikipedia's potential.

Ferrández *et al.* [2007] use Wikipedia for cross-language question answering (see Section 4.3 for research on monolingual question answering). They identify named entities in the query, link them to Wikipedia article titles, and derive equivalent translations in the target language. Wikipedia's exceptional coverage of named entities (Section 3.2.2) counters the main problem of cross-language question answering: low coverage of the vocabulary that links questions to documents in other languages. For example, the question *In which town in Zeeland did Jan Toorop spend several weeks every year between 1903 and 1924?* mentions the entities *Zeeland* and *Jan Toorop*, neither of which is covered by EuroWordNet. In an initial version of the system using that resource, *Zeeland* remains unchanged and *the phrase Jan Toorop* is translated to *Enero Toorop* because *Jan* is erroneously interpreted as *January*. With Wikipedia as a reference, the translation is correct: *¿En qué ciudad de Zelanda pasaba varias semanas al año Jan Toorop entre 1903 y 1924?* With Wikipedia's help, Ferrández *et al.* increase the percentage of correctly answered questions by 20%.

Erdmann *et al.* [2008] show that simply following language links in Wikipedia is insufficient for a high-coverage bilingual dictionary. They develop heuristics based on Wikipedia's link structure that extract significantly more translation pairs, and evaluate them on a manually created test set containing terms of different frequency. Given a Wikipedia article that has been translated into another language—the target article—they augment the translated article name with redirects and also anchor text used to refer to the article. Redirects are weighted by the proportion of links to the target article (including all redirects) that use this particular redirect. Anchors are weighted similarly, by expressing the number of links that use this particular anchor text as a proportion of the total number of incoming links to the article. If a term appears as both redirect and anchor text, the two weights are combined. The resulting dictionary contains all translation pairs whose weight exceeds a certain threshold. This achieves significantly better results than a standard dictionary creation approach using parallel corpora. Figure 5 shows the system in action.

This section has demonstrated Wikipedia's immense potential as a repository of linguistic knowledge for natural language processing. Impressive results have been achieved, particularly on well-defined tasks such as determining semantic relatedness and word sense disambiguation.

4. INFORMATION RETRIEVAL

Given its utility for natural language processing, it is not surprising that Wikipedia has also been used to organize documents and locate them. This section describes

Figure 5. Screen shot of automatically created translations for *plant*.

applications of Wikipedia to information retrieval. These split roughly into *searching* and *browsing*.

For searching, Wikipedia has been leveraged to gain a deeper understanding of both queries and documents, and improve how they are matched to each other. Section 4.1 describes how it has been used to expand queries to allow them to return more relevant documents, while Section 4.2 describes experiments in cross-language retrieval. Wikipedia has also been used to retrieve specific portions of documents, such as answers to questions (Section 4.3) or important topics (Section 4.4).

For browsing, the same Wikipedia-derived understanding has been used to automatically organize documents into helpful groups. Section 4.5 shows how Wikipedia has been applied to document classification, where documents are categorized under broad headings like *Sport* and *Technology*. To a lesser extent it has also been used to determine the main topics that documents discuss, so that they can be organized under more specific tags (Section 4.6).

4.1 Query expansion

Query expansion aims to improve queries by adding terms and phrases, such as synonyms, alternative spellings, and closely related concepts. Such query reformulations can be performed automatically—without the user's input—or interactively—where the system suggests modifications that could be made.

Milne et al. [2007] use Wikipedia to provide both forms of expansion in their knowledge-based search engine Koru.[15] They first obtain a subset of Wikipedia articles that are relevant for a particular document collection, and use the links between these to

[15] Demo at *http://www.nzdl.org/koru*

build a corpus-specific thesaurus. Given a query they map its phrases onto topics in this thesaurus. Figure 6 demonstrates how a query *president bush controversy* is mapped to potentially relevant thesaurus topics (or Wikipedia articles) *George H.W. Bush*, *George W. Bush* and *Controversy. President Bush* is initially disambiguated to the younger of the two, because he occurs most often in the document set. This can be corrected manually. The redirects from his article and that of *Controversy* are then mined for synonyms and alternative spellings, such as *Dubya* and *disagreement*, and quotes are added around multi-word phrases (such as *Bush administration*). This results in a complex Boolean query such as an expert librarian might issue. The knowledge base derived from Wikipedia was capable of recognizing and lending assistance to 95% of the queries issued to it. Evaluation over the TREC HARD Track [Allan 2005] shows that the expanded queries are significantly better than the original ones in terms of overall F-measure.

Milne et al. also provided interactive query expansion by using the detected query topics as starting points for browsing the Wikipedia-derived thesaurus. For example, *George Bush* provides a starting point for locating related topics such as *Dick Cheney*, *Terrorism*, and *President of the United States*. The evaluation of such exploratory search provided little evidence that it assisted users. Despite this, the authors argue that Wikipedia should be an effective base for this task, due to its extensive coverage and inter-linking. This is yet to be proven, however: to our knowledge there are no other examples of exploratory searching with Wikipedia.

Li et al. [2007] also use Wikipedia to expand queries, but focus on the most problematic ones; those that traditional approaches fail to improve. The standard method for improving queries—pseudo-relevance feedback—works by feeding terms from the highest ranked documents back into the query [Ruthven and Lalmas 2003]. This works well in general, so most of the state-of-the-art approaches are variants of this idea. Unfortunately it makes bad queries even worse, because it relies on at least the top few documents being relevant. Li et al. avoid this by using Wikipedia as an external corpus to obtain additional query terms. They issue the query on Wikipedia to retrieve relevant articles. They then use these articles' categories to group them, and rank articles so that those in the largest groups appear more prominently. Forty terms are then picked from the top 20 articles—it is unclear how they are selected—and added to the original query. When tested on queries from TREC's 2005 Robust track [Allan 2005], this improved those queries on which traditional pseudo-relevance feedback performs most poorly. It did not perform as well as the state of the art in general, however. The authors attribute this

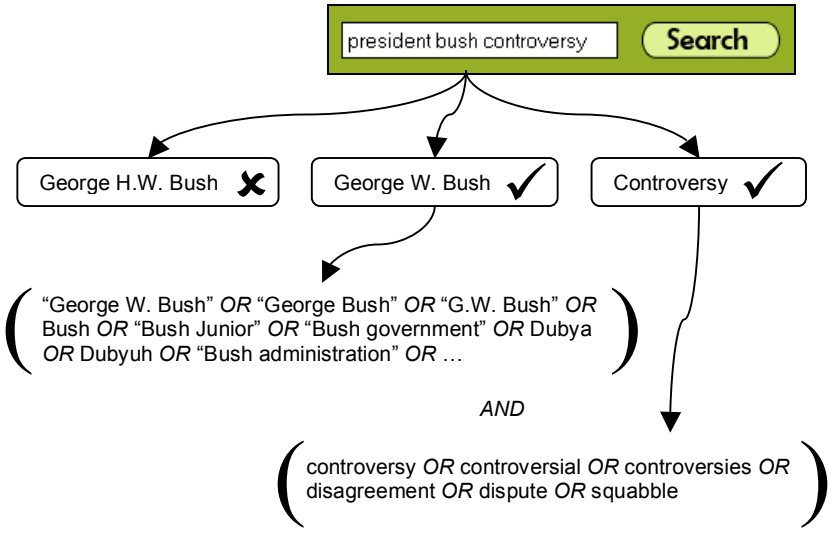

Figure 6. Using Wikipedia to recognize and expand query topics.

to differences in language and context between Wikipedia and the dated news articles used for evaluation, which render many added terms irrelevant.

Where the previous two systems departed from traditional bag-of-words relevance feedback, Egozi et al. [2008] instead aim to augment it. Their system, MORAG, uses Explicit Semantic Analysis (described in Section 3.1) to represent documents and queries as vectors of their most relevant Wikipedia articles. Comparison of document vectors to the query vector results in concept-based relevance scores, which are combined with those given by state-of-the-art retrieval systems, such as Xapian and Okapi. Additionally, both concept-based and bag-of-words scores are computed by segmenting documents into overlapping 50 word subsections (a common strategy), so that the total score of a document is the sum of the score obtained from its best section and its overall content. One complication that this approach must overcome is ESA's tendency to provide features (Wikipedia articles) that are only peripherally related to queries. The query *law enforcement, dogs,* for example, results not just in *police dog* and *cruelty to animals,* but also *contract* and *Louisiana.* To address this, MORAG first ranks documents according to their BOW scores, and then uses the highest and lowest ranking documents to provide positive and negative examples for selecting features. When used to augment the four top performing systems from the TREC-8 competition [Voorhes and Harman 2000] MORAG achieved improvements of between 4% and 15% to Mean Average Precision, depending on the system being augmented.

We were surprised to find only these three papers on using Wikipedia to expand queries, despite the fact that it seems well suited to this task. Bag-of-words based

approaches stand to benefit from Wikipedia's understanding of what the words mean and how they relate to each other. Concept based approaches that draw on traditional knowledge bases could profit just as much from Wikipedia's unmatched breadth. We expect widespread usage of Wikipedia in the future, both for automatic query expansion and exploratory searching, and for both improving existing techniques and supporting entirely new ones.

4.2 Multilingual Retrieval

Multilingual or cross-language information retrieval involves searching for relevant documents that were not written in the same language as the query, which serves the large number of bilingual or multilingual users. Wikipedia has clear application to this task. Although its language versions grow at different rates and cover different topics, they are carefully interwoven. For example, the English article on *Search engines* is linked to the German *Suchmaschine*, the French *Moteur de recherché*, and more than 40 other translations. These links constitute a comprehensive cross-lingual dictionary of topics and terms, which is growing rapidly. This makes Wikipedia ideal for translating emerging named entities and topics, such as people and technologies—exactly the items that more traditional multilingual resources (dictionaries) struggle with. Surprisingly, we failed to locate any papers that use Wikipedia's cross-language links directly to translate query topics.

Instead Potthast et al. [2008] jump directly to a more sophisticated solution that uses Wikipedia to generate a multilingual retrieval model. This is a generalization of traditional monolingual retrieval models—like the vector space model or latent semantic analysis—which assess similarities between documents and fragments of text. Multilingual and cross-language models are capable of identifying similar documents even when they are written in different languages. Potthast et al. take Explicit Semantic Analysis—which, as described in Section 3.1, represents documents by their most relevant Wikipedia concepts—as the starting point for a new model called Cross-language Explicit Semantic Analysis or CL-ESA. This approach depends on the hypothesis that the relevant concepts identified by ESA are essentially language independent, so long as the concepts are sufficiently described in different languages. If there were sufficient overlap between the English and German Wikipedias, for example, then one would get roughly the same list of concepts (and in the same order) from ESA regardless of whether the document being represented, or the concept space it was projected onto, was in English or German. This means that the languages of documents and concept spaces are largely irrelevant, and documents in different languages can be compared without explicit translation.

To evaluate this idea, Potthast et al. conducted several experiments with a bilingual (German/English) set of 3,000 documents. One test was to use articles in one language as queries, to retrieve their direct translations in the other language. When CL-ESA was used to rank all English documents by their similarity to German ones, the explicit translation of the document was consistently ranked highly—it was first 91% of the time, and in the top 10 more than 99% of the time. Another test was to use an English document as a query for the English document set, and its translation as a query for the German one. The two result sets had an average correlation of 72%. These results were obtained with a dimensionality of 10^5; that is, 100,000 bilingual concepts were used to generate the concept spaces. Today, only German and English Wikipedias have this degree of overlap. Results degrade as fewer concepts are used; Potthast et al. found that between 1,000–10,000 concepts are sufficient for reasonable retrieval performance. At the time, this made CL-ESA capable of pairing English with German, French, Polish, Japanese, and Dutch. In time, improvements to the algorithm and continued growth of Wikipedia will allow these techniques to be applied to other languages as well.

4.3 Question answering

Question answering is a more complex form of information retrieval, which aims to return specific answers to questions, rather than entire documents. This ranges in sophistication from merely obtaining the most relevant sentences or sections from documents, to ensuring that they are in the correct form to constitute an answer, to constructing answers on the fly. Wikipedia provides an extremely broad corpus filled with numerous facts, which makes it a promising source of answers. A simple but well-known example of this is how Google queries prefixed with *define*, and Ask.com queries starting with *What is…* or *Who is…*, often return the first sentences from relevant Wikipedia articles.

Kaisser's [2008] QuALiM system, illustrated in Figure 7, provides a more sophisticated example of question answering with Wikipedia.[16] When asked a question (such as *Who is Tom Cruise married to?*) it mines Wikipedia not only for relevant articles, but also for the sentences and paragraphs in which the answer is given. It also provides the exact entity that answers the question—e.g. *Katie Holmes*. Interestingly, this entity is not mined from Wikipedia but obtained by analyzing results from various web search engines. It parses questions to identify the expected class of the answer (in this case, a person), and construct valid queries (e.g. *Tom Cruise is married to* or *Tom Cruise's wife*). Responses to these queries are then parsed to identify entities of the

[16] Demo at *http://demos.inf.ed.ac.uk:8080/qualim/*

correct type to satisfy the answer. Wikipedia is then only used to provide the supporting sentences and paragraphs.

The TREC series of conferences hosts a prominent forum for investigating question answering,[17] The question-answering track provides ground truth for experiments with a corpus from which answers to questions have been manually extracted. The 2004 track saw two of the first uses of Wikipedia for question answering, from Lita et al. [2004] and Ahn et al. [2005]. The former does not perform question answering *per se*; instead it investigates whether different resources provide answers to questions, without attempting to extract the answers automatically. Wikipedia's coverage of answers was 10 percentage points higher than WordNet, and about 30 points higher than the other resources they compared it to, including Google *define* queries and gazetteers such as the CIA *World Fact Book*.

Ahn et al. [2005] seem to be the first to provide explicit answers from Wikipedia. They first identify the topic of the question—*Tom Cruise* in our example—and locate the relevant article. They then identify the expected type of the answer—in this case, another person (his wife)—and scan the article for matching entities. These are ranked by both

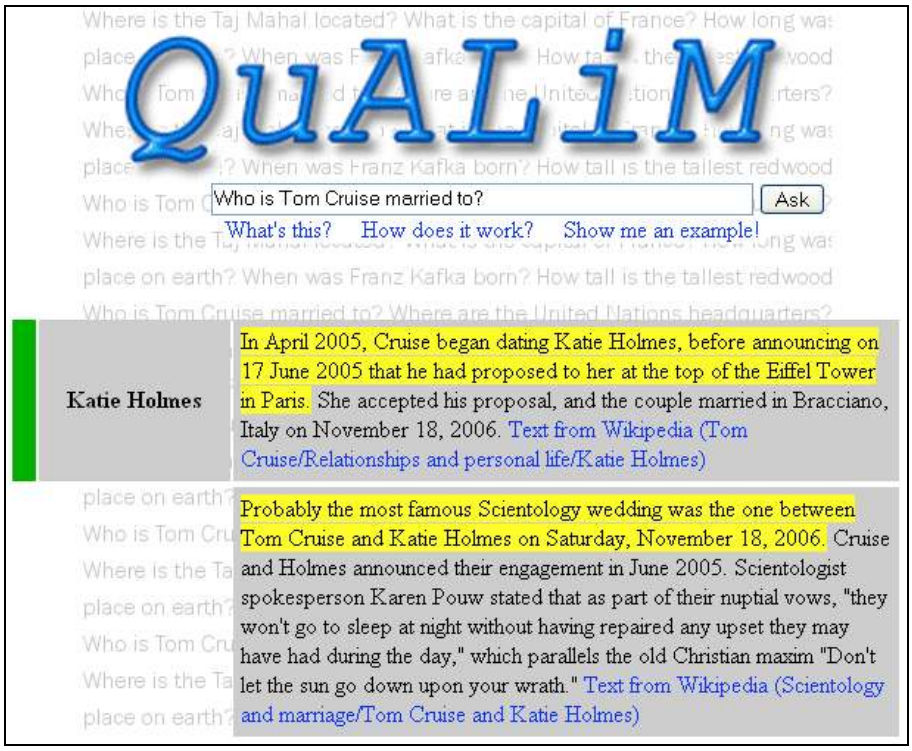

Figure 7. The QuALiM system, using Wikipedia to answer *Who is Tom Cruise married to?*

[17] *http://trec.nist.gov/*

prior answer confidence (probability that they answer any question at all) and posterior confidence (probability that they answer the question at hand). Prior confidence is given by the position of the entity in the article, since articles cover the most important facts first. Posterior confidence is given by the Jaccard similarity of the original question and the sentence surrounding the entity. Wikipedia is used as one stream among many from which to extract answers, and unfortunately the experiments do not tease out its specific contribution. Consequently is difficult to measure the effectiveness of their approach. Overall, however, they describe the results as "disappointing" because it did not improve upon their previous work.

The CLEF series of conferences and competitions is another popular forum for investigating question answering.[18] Monolingual and cross-language QA are addressed by providing corpora and tasks in many different languages. One source of documents is a cross-language crawl of Wikipedia. Most entries for this competition extract answers from Wikipedia but are not covered here because they do not take advantage of its unique properties.

Buscaldi and Rosso [2007a] use Wikipedia to augment their question answering system QUASAR. The way in which this system extracts answers was left unchanged, except for an additional step where Wikipedia is consulted to verify the results. They index four different views of Wikipedia—titles, full text, first sections (definitions), and the categories that articles belong to—and search them differently depending on the question type. Answers to definition questions (e.g., *Who is Nelson Mandela?*) are verified by seeking articles whose title contains the corresponding entity and whose first section contains the proposed answer. If the question requires a name (e.g., *Who is the President of the United States?*) the process is reversed: candidate answers (*Bill Clinton, George Bush*) are sought in the title field and query constraints (*President*, *United States*) in the definition. In either case, if at least one relevant article is returned the answer is verified. This yielded an improvement of 4.5% over the original system, across all question types. Ferrández et al. [2007] also make use of Wikipedia's structure to answer questions, but focus on cross-lingual tasks, where questions are formulated in a language different from that of the documents from which answers are extracted. Their work is described in Section 3.4.

As well as using Wikipedia as a corpus for standard question answering tasks, CLEF has a track (WiQA) specifically designed to assist Wikipedia's contributors. Its aim, given a source article, is to extract new snippets of information from related articles that should be incorporated into it [Jijkoun and de Rijke 2006]. The authors conclude that the

[18] The homepage for the CLEF series of conferences is at *http://www.clef-campaign.org/*

task is difficult but possible, as long as the results are used in a supervised fashion. The best out of seven participating teams added an average of 3.4 perfect (important and novel) snippets to each English article, with a precision of 36%. Buscaldi and Rosso [2007b], one of the contributing entries,[19] search Wikipedia for articles containing the text of the target article's title. They extract snippets from them, rank them according to their similarity to the original article using the standard bag-of-words model, and discard those that are redundant (too similar) or irrelevant (not similar enough). On English data this yields 2.7 perfect snippets per topic, with a precision of 29%. On Spanish data it obtains 1.8 snippets with 23% precision.

Higashinaka et al. [2007] extract questions, answers and even hints from Wikipedia to automatically generate "*Who am I?*" quizzes. The first two tasks are simple because the question is always the same and the answer is always a person. The challenging part is extracting hints (which are essentially facts about the person) and ranking them so that they progress from vague to specific. They used machine learning for this, based on biographical Wikipedia articles whose facts have been manually ranked.

Overall, research on question answering tends to treat Wikipedia as just another plain-text corpus from which to extract answers. Few researchers take advantage of Wikipedia's unique structural properties (e.g. categories, links, etc) or the explicit semantics it provides. Instead they apply standard word-based similarity measures, even when Wikipedia concept-based measures such as ESA have been proven to be more effective. We were surprised to find little overlap between this work and research on information extraction from Wikipedia (Section 5), and no use of Wikipedia derived ontologies or its infoboxes (Section 6). Perhaps this reflects an overall goal of crawling the entire web for answers, requiring techniques that are generalizable to any textual resource.

4.4 Entity ranking

It is often expedient to return entities in response to a query rather than full documents as in classical retrieval. This resembles question answering and often fulfils the same purpose—for example, the query *countries where I can pay in euros* could be answered by a list of relevant countries. For other queries, however, entity ranking does not provide answers but instead generates a list of pertinent topics. For example, as well as *Google*, *Yahoo*, and *Microsoft Live* the query *search engines* would also return *PageRank* and *World Wide Web*. The literature seems to use the term entity and named

[19] We were unable to locate papers describing the others.

entity interchangeably, thus it is unclear whether concepts such as *information retrieval* and *full text search* would also be valid results.

Section 5.3 demonstrates that Wikipedia offers an exceptionally large pool of manually-defined entities, which can be typed (as people, places, events, etc.) fairly accurately. The entity ranking track of the Initiative for Evaluation of XML Retrieval (INEX) compares different methods for entity ranking by how well they are able to return relevant Wikipedia entities in response to queries [de Vries et al. 2007]. Zaragoza et al. [2007] also use Wikipedia as a dataset for comparing two main approaches to entity ranking: entity containment graphs and web search based methods. Their results are of little interest here because they do not relate directly to Wikipedia. More relevant is that they have developed a version of Wikipedia that has been automatically annotated with named entities, and are sharing it so that others can investigate different approaches to named entity ranking.[20]

As well as a being source of entities, Wikipedia provides a wealth of information about them, which can improve ranking. Vercoustre et al. [2008] combine traditional search with Wikipedia-specific features. They rank articles (which they assume are synonymous with entities) by combining the score provided by a search engine (Zettair) with features mined from categories and inter-article links. The article links provide a simplified PageRank for entities and the categories provide a similarity score for how they relate to each other. The resulting precision is almost double that of the search engine alone. Vercoustre et al. were the only competitors for the INEX entity-ranking track we were able to locate,[21] and it seems that Wikipedia's ability to improve entity ranking has yet to be evaluated against more sophisticated baselines. Moreover, the features that Vercoustre et al. derive from Wikipedia are only used to rank entities in general, not by their significance for the query. Regardless, entity ranking will no doubt receive more attention as the INEX competition grows and others use Zaragossa et al.'s dataset.

The knowledge that Wikipedia provides about entities can also be used to organize them. This has not yet been thoroughly investigated, the only example being Yang et al.'s [2007] use of Wikipedia articles and WikiBooks to organize entities into hierarchical topic maps. They search for the most relevant article and book for a query and simply strip away the text to leave lists of links—which again they assume to be entities—under the headings in which they were found. This is both a simplistic entity ranking method and a tool for generating domain-specific taxonomies, but has not been evaluated as either.

---

[20] The annotated version of Wikipedia is at *http://www.yr-bcn.es/semanticWikipedia*
[21] It began in 2007 and the Proceedings are yet to be published.

4.5 Text categorization

Text categorization (or classification) organizes documents into meaningful homogeneous groups. Documents are labeled from a pool of categories in the same way that articles in a newspaper are assigned to sections like business, sport, or entertainment. The traditional approach to this task is to represent documents with the words they contain, and use training documents to identify the words and phrases that are most indicative of each category label. Wikipedia allows categorization techniques to draw on background knowledge about the concepts these words represent. As Gabrilovich and Markovitch [2006] note, traditional approaches are brittle. They break down when documents discuss similar topics in different terms—as when one talks of *Wal-Mart* and the other of *department stores*. They cannot make the necessary connections because they lack background knowledge about what the words mean. Wikipedia can fill the gap.

As a quick indication of Wikipedia's application to text categorization, Table 4 compares Wikipedia-based approaches with state of the art categorization that only uses information obtained from the documents themselves. The figures were obtained on the Reuters-21578 collection, a set of news stories that have been manually assigned to categories. Results are presented as the break even point (BEP) where recall and precision are equal. The micro and macro columns correspond to how these are averaged: the former averages across documents, so that smaller categories are largely ignored; while the latter averages by category. The first entry is a baseline provided by Gabrilovich and Markovitch, which is in line with state-of-the-art document-based methods such as [Dumais et al. 1998]. The remaining three entries use additional information gleaned from Wikipedia and are described below. The gains may seem slight, but they represent the first improvements upon a performance plateau reached by previous state-of-the-art techniques, which are now a decade old.

Gabrilovich and Markovitch [2006] observed that documents can be augmented with Wikipedia concepts without complex natural language processing. Both are in the same form—plain text—so standard similarity algorithms can be used to compare documents with potentially relevant articles. Thus documents can be represented weighted lists of

	Micro BEP	Macro BEP
Baseline (from Gabrilovich and Markovitch [2006])	87.7	60.2
Gabrilovich and Markovitch [2006]	88.0	61.4
Wang et al. [2007]	91.2	63.1
Minier et al. [2007]	86.1	64.1

Table 4. Performance of text categorization over the Reuters-21578 collection.

relevant concepts, rather than bags of words. This should sound familiar; it is the predecessor of Explicit Semantic Analysis, an influential technique that we have seen several times before (Section 3.1, 4.1, 4.2). For each document, Gabrilovich and Markovitch generate a large set of features (articles) not just from the document as a whole, but also by considering each word, sentence, and paragraph independently. Training documents are then used to filter out the best of these features, to augment the original bags of words. Additionally the number of links made to each article is used to identify and emphasize those that are most well known. This results in consistent improvements over the previous classification techniques, particularly over short documents (which otherwise have few features) and small categories (which provide fewer training examples).

The ability of Wikipedia to improve classification of short documents is confirmed by Banerjee et al. [2007], who focus on clustering news articles under feed items such as those provided by Google News. They took a simple approach for obtaining relevant articles for each news story, by issuing its title and short description (Google snippet) as separate queries to a Lucene index of Wikipedia. They were able to cluster the documents under their original headings (each feed item organizes many similar stories) with 90% accuracy using only the titles and descriptions as input. This work is somewhat suspect, however, in that it treats Google's automatically clustered news stories as ground truth, and only compares their Wikipedia-based approach to a baseline of their own design.

Wang et al. [2007] also use Wikipedia to improve document classification, but focus on mining Wikipedia for terms and phrases to add to the bag of words that represent each document. For each document, they locate relevant Wikipedia articles by matching n-grams to article titles. They then augment the document by crawling these articles for synonyms (redirects), hyponyms (parent categories) and associative concepts (inter-article links). In the latter case they acknowledge that many links exist between articles that are only tenuously related at best. They overcome this by only selecting linked articles that are closely related according to textual content or parent categories. As shown in Table 4, this results in the best overall performance.

As well as a source of background knowledge for improving classification techniques, Wikipedia can be used as a corpus for training and evaluating them. Almost all classification approaches are machine-learned, and thus require training examples. Wikipedia provides millions of them. Each association between an article and the categories to which it belongs can be considered as manually defined ground truth for how that article should be classified. Gleim et al. [2007], for example, use it to evaluate their techniques for categorizing web pages solely on their structure rather than textual

content. Admittedly, this is a well-established research area with well-known datasets, so it is unclear why another one is required. Table 4, for example, would be more informative if all of the researchers using Wikipedia for document classification had used standard datasets instead of creating their own.

Two interesting approaches that do not compete with the traditional bag-of-words approaches (and will therefore be discussed only briefly) are Janik and Kochut [2007] and Minier et al. [2007]. The former is one of the few techniques that does not use machine learning for classification. Instead Janik and Kochut mine miniature "ontologies"—rough networks of relevant concepts—from Wikipedia for each document and category, and algorithmically identify the most relevant category ontology for each document ontology. The latter approach transforms the document-term matrix used by traditional techniques by mapping it onto a gigantic term-concept matrix obtained from Wikipedia. PageRank is run over Wikipedia's inter-article links in order to weight the derived Wikipedia concepts, and dimensionality reduction techniques (latent semantic analysis, kernel principle component analysis and kernel canonical correlation analysis) are used to reduce the representation to a manageable size. Minier et al. attribute the disappointing results (shown in Table 4) to differences in language usage between Wikipedia the Reuters corpus used for evaluation. It should be noted that their Macro BEP (the highest in the Table) may be misleading; their baseline achieves an even higher result, indicating that their experiment should not be compared to the other three.

Banerjee [2007] observed that document categorization is a problem where the goalposts shift regularly. The typical application is organizing news stories or emails, which arrive in a constant stream where the topics being discussed constantly evolve. A categorization method trained today may not be particularly helpful next week. Instead of throwing away old classifiers, they show that inductive transfer allows old classifiers to influence new ones. This improves results and reduces the need for fresh training data. They find that classifiers which derive additional knowledge from Wikipedia are more effective at transferring this knowledge, which they attribute to Wikipedia's ability to provide background knowledge about the content of articles, making their representations more stable.

Dakka and Cucerzan [2008] and Bhole et al. [2007] perform the reverse of the above techniques. Instead of using Wikipedia to augment document categorization, they apply categorization techniques to Wikipedia. Their aim is to classify articles to detect the types (people, places, events, etc.) of the named entities they represent. Since this has more to do with named entity recognition than document classification, discussion of it is deferred to Section 5.3. Also discussed elsewhere is Schönhofen [2006] who developed

a topic indexing system but evaluated it as a document classifier. His work is left for the next section.

Overall, the use of Wikipedia for text categorization is a flourishing research area. Many recent efforts have improved upon the previous state of the art; a plateau that had stood for almost a decade. Some of this success may be due to the amount of attention the problem has generated (at least 10 papers in just three years), but more fundamentally it can be attributed to the way in which researchers are approaching the task. Just as we saw in Section 4.1, the greatest gains have come from drawing closely on and augmenting existing research, while thoroughly exploring the unique features that Wikipedia offers.

4.6 Topic Indexing

Topic indexing is subtly different from text categorization. Both label documents so that they can be grouped sensibly and browsed efficiently, but in topic indexing labels are chosen from the topics the documents discuss rather than from a predetermined pool of categories. Topic labels are typically obtained from a domain-specific thesaurus—such as MESH [Lipscomb 2000] for the Medical domain—because general thesauri like WordNet and Roget are too small to provide sufficient detail. An alternative is to obtain labels from the documents themselves, but this is inconsistent and error-prone because topics are difficult to recognize and appear in different surface forms. Using Wikipedia as a source of labels sidesteps the onerous requirement for developing or obtaining relevant thesauri, since it is large and general enough to apply to all domains. It might not achieve the same depth as domain-specific thesauri, but tends to cover the topics that are used for indexing most often [Milne et al. 2006]. It is also more consistent than extracting terms from the documents themselves, since each concept in Wikipedia is represented by a single succinct manually chosen title. In addition to the labels themselves, Wikipedia provides many additional features about the concepts, such as how important and well known they are, and how they relate to each other.

Medelyan et al. [2008] propose topic indexing that uses Wikipedia as a controlled vocabulary and applies wikification (defined in Section 3.2.1) to identify the topics mentioned within documents. For each candidate topic they identify several features, including classical, such as how often topics are mentioned, and two Wikipedia-specific ones. One is *node degree*: the extent to which each candidate topic (article) is linked to the other topics detected in the document. The other is *keyphraseness*: the extent to which the topics are used as links in Wikipedia. They use a supervised approach that learns the typical distributions of these features from manually tagged corpus [Frank et al. 1999]. For training and evaluation they had 30 people, working in pairs, index 20

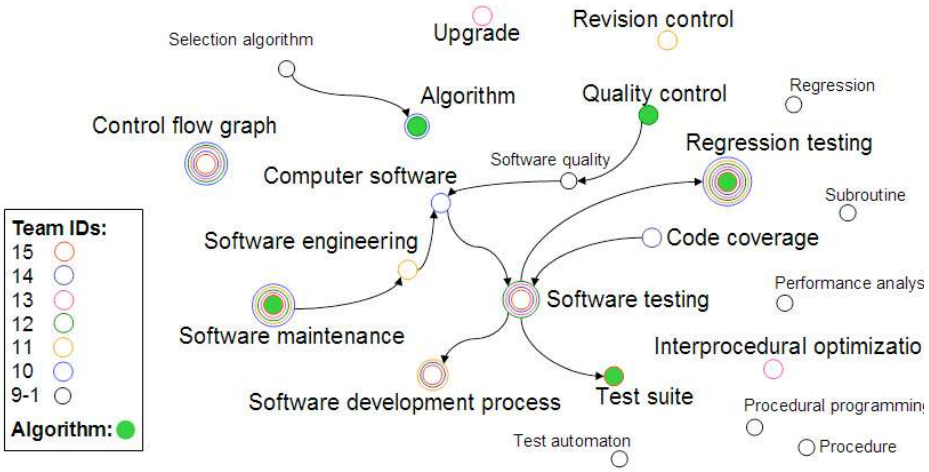

Figure 8. Topics assigned to a document entitled "A Safe, Efficient Regression Test Selection Technique" by human teams (outlined circles) and the new algorithm (filled circles).

documents. Figure 8 shows key topics for one document and demonstrates the inherent subjectivity of the task—the indexers did not all choose the same topics, and achieved only 30% agreement with each other. Medelyan et al.'s automatic system, whose choices are shown as filled circles in the figure, obtained the same level of agreement and requires little training.

Although it has not been evaluated as such, Gabrilovich and Markovitch's [2007] Explicit Semantic Analysis, described in Section 3.1, essentially performs topic indexing. For each document or text fragment it generates a weighted list of relevant Wikipedia concepts, the strongest of which should be suitable topic labels. Another approach that has not been compared to manually indexed documents is Schönhofen [2006], who uses Wikipedia categories as the vocabulary from which key topics are selected. Documents are scanned to identify the article titles and redirects they mention, and documents are represented by the categories that contain these articles—weighted by how often the document mentions the category title, its child article titles, and the individual words in them. Schönhofen did not compare the resulting categories with index topics, but instead used them to perform document categorization. Roughly the same results are achieved whether documents are represented by these categories or by their content in the standard way. Combining the two yields a significant improvement.

Like document categorization, research in topic indexing builds solidly on related work, but has been augmented to make interesting use of Wikipedia. Although not a great deal of research has been done, significant gains have been achieved over the previous state of the art. The results have not yet been evaluated as rigorously as in categorization, however. Medelyan et al. [2008] have directly compared their results

against manually defined ground truth, but this was restricted to a relatively small dataset. To advance further, larger datasets need to be developed for evaluation and training.

5. INFORMATION EXTRACTION

Where information retrieval is driven largely by the goal of answering specific questions, information extraction seeks to deduce meaningful structures from unstructured data such as natural language text, though in practice the dividing line between the fields is not sharp. These structures are usually represented as relations. For example, from this:

> *Apple Inc.'s world corporate headquarters are located in the middle of Silicon Valley, at 1 Infinite Loop, Cupertino, California.*

a relation *hasHeadquarters*(*Apple Inc., 1 Infinite Loop-Cupertino-California*) might be extracted. The challenge is to extract this relation from sentences expressing the same information about *Apple Inc.*, regardless of the actual wording. Moreover, given a similar sentence about other companies, the same relation should be determined with different arguments, e.g., *hasHeadquarters*(*Google Inc., Google Campus-Mountain View-California*).

Methods for extracting relations from Wikipedia can be grouped into those that use its raw text (Section 2.3.2) and those that use its semi-structured parts and internal hyperlink structure (Section 2.3.3, 2.3.4 and 2.3.5). The former, described in Section 5.1, apply methods developed before Wikipedia was recognized as a linguistic resource; for them, any text represents a source of relations. The extraction process benefits from the encyclopedic nature of Wikipedia articles and their uniform writing style. The latter, described in Section 5.2, exploit unique Wikipedia properties such as infoboxes and the category structure. Finally, in Section 5.3 the determination of named entities and their type is treated as a task of its own. As noted earlier, Wikipedia's coverage of named entities is uniquely comprehensive and up-to-date (Section 3.2.3). Such work extracts named entity information such as *isA*(*Portugal, Location*) and *isA*(*Bob Marley, Person*). Again, although the task is similar to that in Sections 5.1 and 5.2, different techniques are applied, like analysis of geographical coordinates.

5.1 Semantic relations in Wikipedia's raw text

Extracting semantic relations from raw text begins by taking known relations that serve as seeds and extracting patterns from their text—*X's * headquarters are located in * at Y* in the above example. These patterns are applied to a large text corpus to identify new relations. For this, a phrase chunker or named entity recognizer is applied to identify entities that appear in a sentence, intervening patterns are compared to the seed patterns,

Figure 9. Wikipedia's description of the *Waikato River*.

and when they match, new semantic relations are discovered. Culotta *et al.* [2006] summarize difficulties in this process:

- Enumerating over all pairs of entities yields a low density of correct relations even when restricted to a single sentence
- Errors in the entity recognition stage create inaccuracies in relation classification.

Wikipedia's structure helps combat these difficulties. Each article represents a particular concept that serves as a clearly recognizable *principal entity* for relation extraction from that article. Its description contains links to other, secondary, entities. All that remains is to determine the semantic relation between these entities. For example, the description of the *Waikato River*, shown in Figure 9, links to entities like *river*, *New Zealand, Lake Taupo* and many others. Appropriate syntactic and lexical patterns can extract a host of semantic relations between these items.

Ruiz-Casado *et al.* [2005] mine relations from Simple Wikipedia using WordNet as a source of positive examples (Ruiz-Casado *et al.* [2007] explain the technique in greater detail). Given two co-occurring semantically related WordNet nouns in a Wikipedia article, the intervening text is used to find relations that are absent from WordNet. But first the text is generalized. If the edit distance falls below a predefined threshold—i.e., the two strings nearly match—those parts that do not match are replaced by a wildcard (*). For example, a generalized pattern: *X directed the * famous|known film Y* is obtained from two strings: *X directed the famous film Y* and *X directed the well known film Y*. Using this technique Ruiz-Casado *et al.* identify 1200 new semantic relations with a precision of 61–69% depending on the relation type.

Ruiz-Casado *et al.* [2006] generalize this technique to extract relations between automatically identified entities without using WordNet as a reference. The English

Wikipedia is used as a corpus, but now the authors concentrate only on those parts that are likely to contain relations of interest. They crawl Wikipedia's list pages to access *prime ministers*, *authors*, *actors*, *football players*, and *capitals*; and infer the same kind of predefined patterns as above. They manually evaluate precision on at least 50 examples for each relation type. If the pages are combined into a single corpus results vary wildly, from 8% precision on the *player-team* relation to 90% for *death-year*. The reason is heterogeneity in style and mark-up of articles. When the *player-team* patterns are applied just to articles about football players, precision increases to 93%.

Herbelot and Copestake [2006] extract hyponymy relations from sentences containing the verb *to be* (including *is*, *was*, *will be* etc.) Instead of performing simple pattern matching of the form *X is a Y* with some wildcards, they analyze the sentences to identify the subject, object and their relationship, regardless of word order. These authors use their own dependency analyzer, called *Robust Minimal Recursion Semantics*, which can handle partially parsed sentences. This analyzer re-organizes a parsed sentence into a series of minimal semantic trees whose root elements correspond to lemmas in the sentence. The same tree is obtained for similar sentences like *Xanthidae is a family of crabs* and *Xanthidae is one of the families of crabs* (Figure 10).

The results are evaluated manually on a subset of 100 articles and automatically using a thesaurus, restricted it to Wikipedia articles describing animal species. Because only 3 patterns were used, recall was low: 14% at precision 92%. To improve recall they suggest extracting patterns automatically. The same dependency analyzer is used, which yields patterns that are more general than regular expressions, although no explicit performance comparison is provided. Initial experiments increase recall to 37%; however, precision drops to 65%.

Suchanek *et al.* [2006] also employ linguistic techniques to achieve better results than regular expressions. They parse each sentence with a context-free grammar. A pattern is defined by a set of syntactic links between two given concepts, called a *bridge*. For example, the bridge in Figure 11 matches sentences like *Chopin was great among the composers of his time* where *Chopin=X* and *composers=Y*. Machine learning techniques are applied to determine and generalize patterns that describe relations of interest from manually supplied positive and negative examples. The approach is evaluated on article sets with different degrees of heterogeneity: articles about composers, geography, and random articles. As expected, the more heterogeneous the corpus the worse the results, with best results achieved on composers for the relations *birthDate* (F-measure 75%) and *instanceOf* (F-measure 79%). Unlike Herbelot and Copestake [2006], Suchanek *et al.* show that their approach outperforms other systems, including a shallow pattern

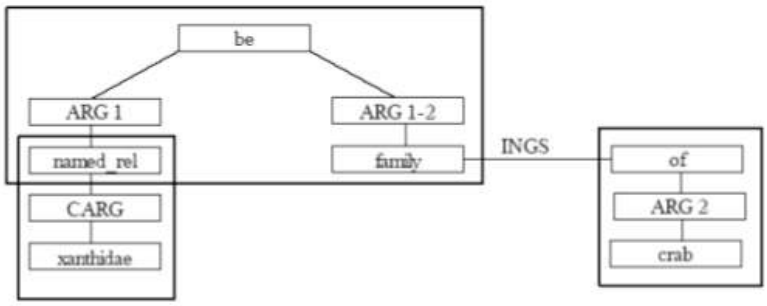

Figure 10. Output of the Robust Minimal Recursion Semantics analyzer for the sentence *Xanthidae is one of the families of crabs* [Herbelot and Copestake, 2006].

matching resource TextToOnto[22] and the more sophisticated scheme of Chimiano and Volker [1995].

Nguyen *et al.* [2007a, 2007b] augment these ways of combining lexical and syntactic patterns with techniques such as anaphora resolution (to increase coverage), full dependency parsing and subtree mining. Sentences are analyzed with OpenNLP[23] and anaphora and co-referents resolved using a simple heuristic developed specially for the purpose. Thus, in an article about the software company 3PAR, phrases like *3PAR*, *manufacturer*, *it* and *company* are tagged as the same principal entity. Next, all link anchors in the article are tagged as secondary entities—ones relating to the principal entity. Sentences with at least one principal and one secondary entity are analyzed by the Minipar dependency parser. The dependency tree of Figure 12a is extracted from the sentence *David Scott joined 3PAR as CEO in January 2001* and is then generalized to match similar sentences (Figure 12b). The subtrees are extracted from a set of training sentences containing positive examples and then applied as patterns to find new semantic relations. The scheme was evaluated using 3,300 manually annotated entities, 200 of which were reserved for testing. 6,000 Wikipedia articles, including 45 test articles, were used as the corpus. The new approach achieved an F-measure of 38%, with precision significantly higher than recall, significantly outperforming two simple baselines.

Wang *et al.* [2007a] use selectional constraints in order to increase the precision of regular expressions without reducing coverage. They also automatically extract positive seeds from infoboxes. For example, the infobox field *Directed by* describes relation *hasDirector*(*FILM, DIRECTOR*) with positive examples <*Titanic, James Cameron*> and

[22] *http://sourceforge.net/projects/texttoonto*
[23] *http://opennlp.sourceforge.net/*

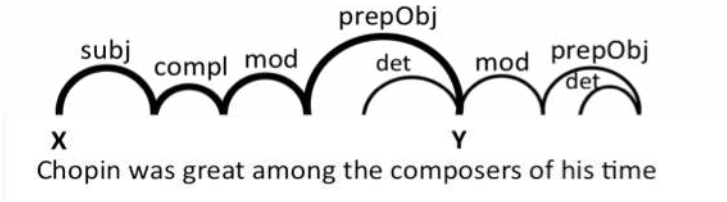

Figure 11. Example bridge pattern used in Suchanek et al. [2006].

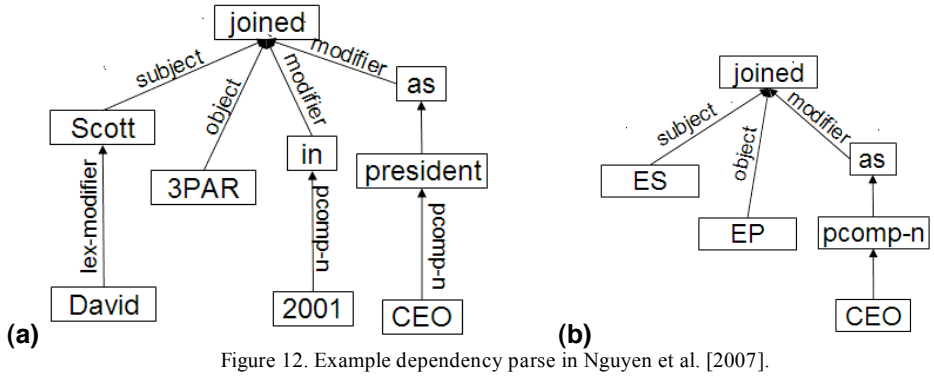

Figure 12. Example dependency parse in Nguyen et al. [2007].

<*King Kong* (*2005*), *Peter Jackson*>. They collect patterns that intervene between these entities in Wikipedia's text and generalize them into regular expressions like

X (*is*|*was*) (*a*|*an*) * (*film*|*movie*) *directed by Y*.

Selectional constraints restrict the types of subject and object that can co-occur within such patterns. For example, Y in the pattern above must be a *director*—or at least a *person*. The labels specifying the types of entities implemented as features are derived using words commonly occurring in Wikipedia articles describing these entities. For example, instances of *ARTIST* extracted from a relation *hasArtist*(*ALBUM, ARTIST*) often co-occur with terms like *singer*, *musician*, *guitarist*, *rapper*, etc. To ensure better coverage, Wang *et al.* cluster such terms hierarchically. The advantage of selectional constraints is that they allow patterns such as 'X's Y' and 'X of Y' to be applied.

The relations *hasDirector* and *hasArtist* are evaluated independently on a sample of 100 relations extracted automatically from the entire Wikipedia and were manually assessed by three human subjects. An unsupervised learning algorithm was applied, and the features were tested individually and together. The authors report precision and accuracy values close to 100%.

The same authors investigate a different technique that does not rely on patterns at all [Wang *et al.* 2007b]. Instead, features are extracted from two articles before determining their relation:

- The first noun phrase and its lexical head that follows the verb *to be* in the article's first sentence (e.g., *comedy film* and *film* in *Annie Hall is a romantic comedy film*)
- Noun phrases that appear in the corresponding category titles and the lexical heads.
- Infobox predicates, e.g. *Directed by* and *Produced by* in *Annie Hall*.
- Terms that appear between the articles in sentences that contain them both as a link.

For each pair of articles the distribution of values of these features is compared with that of positive examples. Unlike in [Wang *et al.* 2007a], no negative instances are used. A special learning algorithm (B-POL) designed for situations where only positive examples are available is applied. First, negative examples are identified from unlabeled data using a weak classifier, and then a strong classifier (e.g., SVM) is used to iteratively classify negative examples until none remain. Four relations were used for evaluation, *hasArtist*(*ALBUM, ARTIST*), *hasDirector*(*FILM, DIRECTOR*), *isLocatedIn*(*UNIVERSITY, CITY*), *isMemberOf*(*ARTIST, BAND*), along with 1,000 named entity pairs classified by three human subjects. Best results were an F-measure of 80% on the *hasArtist* relation, which had the largest training set; the worse was 50% on *isMemberOf*.

Wu and Weld [2007] view the extraction problem as a task of improving infoboxes in Wikipedia. Like Wang *et al.* [2007a, 2007b] they use their content as training data. Their system called Kylin first maps infobox attribute-value pairs to sentences in corresponding Wikipedia article using some simple heuristics. Next, for each attribute it creates a sentence classifier that uses sentence's tokens and their part of speech tags as features. Given an unseen Wikipedia article, a document classifier analyzes its categories and assigns an infobox class, e.g. 'U.S. counties'. Next, sentence classifier is applied to assign relevant infobox attributes. Extracting values from the sentences is treated as a sequential data-labelling problem and Conditional Random Fields are applied for this. Precision and recall of Kylin are measured by its ability to generate correct infoboxes for Wikipedia articles, for which infobox information is known. The authors judged manually the attributes produces by their system and by Wikipedia authors. Kylin's precision ranged from 74 to 97%, at recall levels of 60 to 96% respectively, depending on the infobox class. The authors' precision was around 95% on average and more stable across the classes; their recall was significantly better on most classes but worse or same on others.

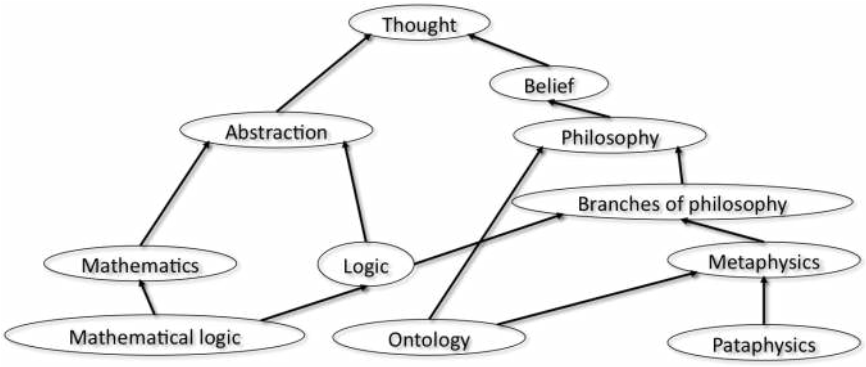

Figure 13. Fragment of Wikipedia's category structure [Ponzetto, 2007].

In a later work Wu et al. [2008] address problems in their approach in the following way. To generate complete infobox schemata for articles of rare classes, they refer to WordNet's ontology and aggregate attributes from parents to their children classes. E.g. knowing that *isA*(*Performer, Person*), infobox for *Performers* receives prior missing field *BirthPlace*. To provide additional positive examples, they apply TextRunner [Banko *et al.* 2007] to the web, in order to retrieve additional sentences describing the same attribute-values pairs. Given a new entity for which an infobox needs to be generated, they use Google search to retrieve additional sentences describing this entity. The combination of these techniques improves the recall by 2 to 9 percentage points while maintaining of increasing precision. Kylin's results are the most complete and impressive in this group of approaches.

The majority presented approaches take advantage of Wikipedia's encyclopedic nature using it as a corpus for extracting semantic relations. Simple pattern matching techniques are outperformed by those that use parsing [Suchanek *et al.* 2006], selectional constraints [Wang *et al.* 2007a] and lexical features [Wang *et al.* 2007b]. Wang *et al.* [2007a] and Wu et al. [2007] show that Wikipedia infoboxes contain positive examples that can improve the extraction if machine learning is applied. Wu et al. [2008] prove that retrieving additional content from the web boosts the extraction performance.

It would be helpful to directly compare the approaches on the same data set. Of course for this, the researchers would need to reach a consensus on what relations they will extract. At this point, while there is an overlap in some relations (*isMemberOf*, *InstanceOf*, *hasDirector*), the choice of a particular relation set by a research group seems to be arbitrary. Furthermore, none of these techniques take advantage of Wikipedia's structural information like hyperlinks between the articles and their categorization. As the next section shows, such information contains a wealth of semantic relations outnumbering the ones appearing in Wikipedia's actual text.

5.2 Semantic relations in structured parts of Wikipedia

Here we describe research that addresses the limitations just identified by seeking semantic relations in (semi-)structured parts of Wikipedia, with the goal of building an alternative to manually created knowledge bases such as WordNet and Cyc. Some label existing links between categories and articles, a process sometimes referred as *link-typing*. As noted in Section 2.2.6, Wikipedia's category structure is made up of what are in fact rather different kinds of relations. For example, in Figure 13 *Category:Mathematical logic* belongs to both *Category:Logic* and *Category:Mathematics*, the former relation should arguably be *isA* and the latter *partOf*. Further differentiation between category relations in Wikipedia is required to transform it into a lexical knowledge base like those created by humans. Some approaches use Wikipedia's infoboxes (Figure 14) as a further source of relational information.

Chernov *et al*. [2006] were one of the first to analyze links between Wikipedia categories. Their goal was to determine semantically strong links, as opposite to "irregular and navigational links." They develop two measures. One correlates semantic strength with the number of hyperlinks between articles assigned to two categories in question; the other is the *connectivity ratio*—the number of links from articles in one category to articles in the other, expressed as a proportion of the total number of links in the first category. Evaluation uses a sample of 100 category pairs, each assessed by human subjects as strongly, averagely or weakly related. Chernov *et al*. observe that both measures correlate with human judgments, but a more thorough study is required.

Several projects extract relations from Wikipedia of a quantity or organization that might properly be called 'ontological'. Discussion of these projects impinges on the territory of Section 6. Here we discuss the projects' methods and relationship to other IE research, while in Section 6 we discuss their end-products considered as ontologies in their own right. One such project is YAGO, *Yet Another Great Ontology* [Suchanek *et al*. 2007]. Here Wikipedia's leaf categories are mapped onto the WordNet taxonomy of synsets, and the articles belonging to those categories are added to the taxonomy as new elements. To perform the mapping, each category's lexical head is extracted—*people* in *Category:American people in Japan* and, if necessary, expressed in singular form—*person*—before being sought in WordNet. If there is a match, it is chosen as the class for this category. This scheme extracts 143,000 *isA* relations—in this case, *isA*(*American people in Japan, person/human*). If more than one match is possible, word sense disambiguation is required (cf. Section 3.2.3). The authors experimented with mapping a category's subcategories to WordNet and choosing the sense that corresponds to the smallest resulting taxonomic graph. However, they claim that this semantically enhanced

technique does not perform as well as choosing the most frequent WordNet synset for a given term (the frequency values are provided by WordNet), an observation that seems inconsistent with findings by other authors [e.g. Medelyan and Milne 2008] who show that the most frequent sense is not necessarily the intended one (Section 3.2.3).

Having established a large core taxonomy, the authors define a mixed suite of heuristics for extracting further relations to add to it. For instance a name parser is applied to all personal names to identify given and family names, adding 440,000 relations like *familyNameOf*(*Albert Einstein, "Einstein"*). Many heuristics make use of the Wikipedia category names, allowing extraction of relations like *bornInYear*, *establishedIn*, *locatedIn* and others. For example, subcategories of categories ending with *birth* (e.g., *1879 birth*) and *establishments*, correspond to the first two relations. A category like *Cities in Germany* indicates the *locatedIn* relation. This yields 370,000 non-hierarchical, non-synonymous relations. Manual evaluation of sample facts by human judges shows 91–99% accuracy, depending on the relation. Also added are 2M synonymy relations generated from redirects, 40M context relations generated from cross-links between articles, and 2M type relations between categories considered as classes and their articles considered as entities. Section 6.6 discusses the number and kinds of facts in YAGO in more detail, as well as further specifically ontological features, such as its purpose-built ontology language.[24]

Another extremely large-scale relation-extraction project is DBPedia [Auer and Lehmann 2007]. This project analyses Wikipedia's infoboxes and transforms their content into RDF triples. Figure 14 shows part of the infobox from the *New Zealand* article; on the right is the Wiki mark-up used to create it. Extracting information from infoboxes is by no means trivial. The information they contain is expressed in an attribute-value notion, which is rendered inside a wiki page by means of an associated template. There are many different templates, with a great deal of redundancy between them—for example, Auer and Lehmann report separate templates for *Infobox_film*, *Infobox Film*, and *Infobox film*. Recursive regular expressions are used to parse relational triples from all templates that are commonly used in Wikipedia and contain at least several predicates. For example, the *country* template encodes relations like *hasCapital*(*New Zealand, Wellington*) or *hasPrimeMinister*(*New Zealand, Helen Clark*). The templates are taken at face value; no heuristics are applied to verify their accuracy. The URL of each entity linked to from an article is recorded as a unique identifier.

Wikipedia categories are treated as classes and articles as individuals. However, Auer and Lehmann do not say what happens to articles that have corresponding categories, like

{{ Infobox Country or territory |

native_name = New Zealand |
…
capital = [[Wellington]] |

latd = 41 | latm = 17 | latNS = S |
longd = 174 | longm = 27 | longEW = E |

largest_city = [[Auckland]] |

official_languages =
[[New Zealand English|English]] (98%)
[[Māori language|Māori]] (4.2%)
[[New Zealand Sign Language|NZ Sign Language]] (0.6%) |

demonym = [[New Zealand People|New Zealander]],[[Kiwi (people)|Kiwi]] |

government_type =
[[Parliamentary democracy]] and
[[Constitutional monarchy]]
…}}

Figure 14. Wikipedia infobox on New Zealand.

New Zealand; presumably article and category receive different identifiers. Unlike YAGO there is no attempt to place facts in the framework of an overall taxonomic structure of concepts. Apart from the infobox relations, links between categories are merely extracted and labeled with the relation *isRelatedTo*.

The resulting DBPedia dataset contains 115,000 classes and 650,000 individuals sharing 8,000 types of semantic relations. A total of 103M triples are extracted, far surpassing any other scheme.[25] However, 60% of these are internal links derived from Wikipedia's link structure; only 15% are taken directly from infoboxes. Also since there is no evaluation it is difficult to judge how accurate the triples are. Unlike other approaches, DBPedia relies on the accuracy of Wikipedia's contributors, and Auer and Lehmann suggest guidelines for authors in order to improve the quality of infoboxes with time. Section 6.6 further discusses DBPedia in the context of YAGO and other ontologies

Work at the European Media Lab Research (EMLR) takes up the challenge of further differentiating category links independently of the DBpedia project. Ponzetto and Strube

[24] YAGO can be queried online or downloaded from *http://www.mpi-inf.mpg.de/~suchanek/downloads/yago/*

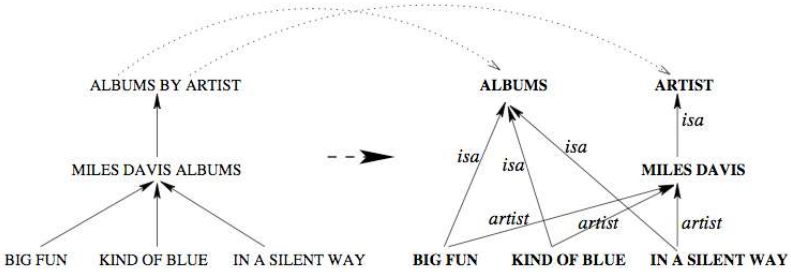

Figure 15. Relations inferred from BY categories [Nastase and Strube 2008].

[2007] observe that the first task is to construct a knowledge taxonomy, or subsumption hierarchy, and that the quickest way to do this is to identify and isolate *isA* relations from amongst already-existing category links. Here *isA* is thought of as subsuming relations between two classes—*isSubclassOf*(*Apples, Fruit*)—and between an instance and its class—*isInstanceOf*(*New Zealand, Country*). They analyze category titles and their connectivity to distinguish between *isA* and what they call '*notIsA'* relations. Several steps are applied in order of accuracy. One of the most accurate matches the lexical head and modifier of two phrases. Sharing the same lexical head indicates an *isA* relation, e.g., *isA*(*British computer scientist, Computer scientist*). Modifier matching indicates *notIsA*, e.g., *notIsA*(*Islamic mysticism, Islam*). Another method uses co-occurrence statistics of two categories within patterns to indicate hierarchical and non-hierarchical relations, e.g., NP_2,*?* (*such as*|*like*|, *especially*) *NP** NP_1 indicates *isA*, and NP_1 *are? used in* NP_2 indicates *notIsA*. This technique induces 100,000 *isA* relations from Wikipedia.

Comparing the derived labels with relations assigned (by knowledge engineers) to concepts with the same lexical heads in ResearchCyc shows that their labeling is highly accurate, depending on the method used, and yields an overall F-measure of 88%. Ponzetto [2007] describes how they plan to apply the induced knowledge base to natural language processing tasks such as co-reference resolution.

Since then the same research group has further refined semantic relations between Wikipedia categories. Zirn *et al.* [2008] divide the derived *isA* relations into those expressing *isSubclassOf* and *isInstanceOf*. For example, *Category:American scientist* generalizes *Category:American physicists*, whereas *Category:Albert Einstein* is an instance of *Category:American physicists*. Two methods assume that all named entities are instances and thus related to their categories by *isInstanceOf*. One uses a named entity recognizer, the other a heuristic based on capitalization in the category title. Further methods include heuristics like: If a category has at least one hyponym that has at least

[25] Further information, and the extracted data, can be downloaded from *http://www.dbpedia.org*

two hyponyms, it is a class. Evaluation against 8,000 categories listed in ResearchCyc as individuals (instances) and collections (classes) shows that the capitalization method is best, achieving 83% accuracy; however, combining all methods into a single voting scheme improves this to 86%. The taxonomy derived from this work is available in RDF Schema format.[26]

Nastase and Strube [2008] extract non-taxonomical relations from Wikipedia by parsing category titles. They are no longer just working with the category network but also deriving entirely new relations between categories, articles and terms extracted from category titles. Explicit unitary relations are extracted—for example, analysis of the category title *Queen* (*band*) *members* results in the *memberOf* relation being inferred from the articles in that category to the article for the band, e.g. *memberOf*(*Brian May, Queen* (*band*)). Explicit binary relations are also extracted—for example, if a category title matches the pattern *X* [*VBN IN*] *Y*, for instance *Movies directed by Woody Allen*, the verb phrase is used to 'type' a relation between all articles assigned to the category and the entity *Y*, e.g. *directedBy*(*Annie Hall*, *Woody Allen*), while the class *X* is used to further type the articles in the category, e.g. *isA*(*Annie Hall, Movie*).

Particularly sophisticated is their derivation of entirely implicit relations from the very common *X by Y* pattern in Wikipedia category names, which facets a great deal of the category structure (e.g. *Writers By Nationality, Writers by Genre, Writers by Language*). For instance, given the category title *Albums By Artist,* they not only label all the articles in the category *isA*(*X, Album*), but also find subcategories pertaining to particular artists (e.g. *MilesDavis, Albums*), locate the article corresponding to the artist, label the entity as an artist, e.g. *isA*(*MilesDavis, Artist*) and label all members of the subcategory as being produced by him, e.g. *artist*(*KindOfBlue MilesDavis*). Figure 15 illustrates this.

Nastase and Strube identify a total of 3.4 million *isA* and 3.2 million *spatial* relations, along with 43,000 *memberOf* relations and 44,000 other relations such as *causedBy* and *writtenBy*. Evaluation with ResearchCyc was not meaningful because of little overlap in extracted concepts—particularly named entities. Instead, human annotators analyzed four samples of 250 relations from the above sets; precision ranged from 84 to 98% depending on relation type. Once again the implications of this work for ontology building will be discussed in Section 6.6.

Although the three approaches presented in this section—YAGO, DBPedia and EMLR's taxonomy—have the same goal, to create an extensive, accurate knowledge base of human language, the techniques differ significantly. The first combines Wikipedia's

[26] *http://www.eml-r.org/english/research/nlp/download/wikitaxonomy.php*

leaf categories (and their instances) with Wordnet's hypernym hierarchy, embellishing this structure with further relations; the second basically dumps the contents of Wikipedia's infoboxes with little further analysis; and the third performs a differentiation or 'typing' of category links, followed by an analysis of category titles and the articles contained by those categories to derive further relations. As a result, the information extracted varies. For instance whereas Suchanek *et al.* [2007] extracts the relation *writtenInYear*, Nastavi and Strube [2008] detect *writtenBy* and Auer and Lehmann [2007] generate *written*, *writtenBy*, *writer*, *writers*, *writerName*, *coWriters*, as well as their case variants. There has so far been little comparison of these approaches, testing of them against each other or attempts to integrate them. We look forward to further research in this area.

5.3 Typing Wikipedia's named entities

One main disadvantage of Wikipedia is its lack of semantic annotation. Infoboxes for entities of the same kind share similar characteristics—for example, *Apple Inc*, *Microsoft* and *Google* share the fields *Founded*, *Headquarters*, *Key People* and *Products*—but Wikipedia does not state that they belong to the same type of named entity, namely *company*. Knowing the type of entity—e.g., *location* or *person*—would supply information that is important for tasks such as information retrieval and question answering (Section 4). This section covers research that classifies articles into predefined classes representing entity-types. The results are semantic relations of a particular kind, e.g. *isA*(*London, Location*).

Toral and Muños [2006] extract named entities from the Simple Wikipedia using WordNet's noun hierarchy. Given an entry—*Portugal*—they extract the first sentence of its definition—*Portugal is a country in the south-west of Europe*—and tag each word with its part of speech. They assign nouns their first (i.e. most common) sense from WordNet and move up in the hierarchy to determine its class, e.g., *country* → *location*. The majority class appearing in the sentence determines the class of the article itself (i.e. entity). The authors achieve 78% F-measure on 404 *locations* and 68% on 236 *persons*. They do not use Wikipedia's special features but mention this as future work.

Buscaldi and Rosso [2007] pursue the same task, but concentrate on locations. Unlike Toral and Muños [2006], they analyze not merely the first sentence but the entire description of each article. In order to determine whether it describes a geographical location, they compare its content with a set of keywords extracted from glosses of locations in WordNet using the Dice metric and cosine coefficient; they also use a multinominal Naïve Bayes classifier trained on the Wikipedia XML corpus [Denoyer and Gallinari 2006]. When evaluated on data provided by Overell and Rüger [2007]

(described in Section 3.2.2) they find that cosine similarity outperforms both the WordNet-based Dice metric and Naïve Bayes, achieving an F-measure of 53% on full articles and 65% on the first sentence. However, the authors fail to achieve Overell and Rüger's [2006] results, and conclude that the content of articles describing locations is less discriminative than other features like geographical coordinates.

Section 3.2.2 discussed how Overell and Rüger [2006, 2007] analyze named entities representing geographic locations, thereby mapping articles to place names listed in a gazetteer. It also described another group of approaches that recognize named entities appearing in raw text and map them to articles. Apart from these, little research has been done on determining the semantic types of named entities. It is surprising that both techniques described in the present section use WordNet as a reference for the entities' semantic class instead of referring to Wikipedia's categories. For example, the three companies mentioned above belong to subcategories of *Category:Companies* and *Portugal* is listed under *Category:Countries*. Moreover, neither technique utilizes the shared infobox fields mentioned above. Annotating Wikipedia with entity labels seems to be low-hanging fruit and we expect to see more advances in the near future.

Approaches to information extraction are less well defined than for natural language processing and most information retrieval tasks, and vary in their scope and depth depending on the research group. There is a dearth of commonly used ground truth data, each technique being evaluated in a different way. It seems that a unified comprehensive general-purpose ontology would be the ideal extension of the research discussed above. For instance, it could unify the specific relations concerning football players and their birth dates extracted from article text with the wealth of taxonomic relations in Wikipedia's category structure and any available named entity information. Thus the next section reviews some of the projects described above, and others, from the perspective of classical, large-scale ontology building.

6. ONTOLOGY BUILDING AND THE SEMANTIC WEB

We now turn to the use of Wikipedia for creating ontologies: comprehensive, large-scale information resources. Section 5 also covers aggregation of knowledge into forms structured for automated reasoning. Nevertheless it is worth treating the topics separately, because ontology building aims for a resource with a level of internal organization and consistency not always found in information extraction. Hence while Section 5 describes the many different methods used for the task, here we consider research projects from the perspective of the comprehensiveness and sophistication of their results, and also the extent to which they contribute to the broad-ranging and ambitious research project known as the semantic web.

6.1 Background: What is Ontology?

A formal ontology is a machine-readable theory of the meanings of some set of concepts or "categories." Building such a resource involves naming the concepts, representing and often categorizing the links between them, and usually encoding some key facts about them. Thus it is generally thought that an ontology which includes the concept *tree* should i) name it as a first-class object (to which synonyms such as the French *arbre* may be attached), ii) link it to closely-related concepts such as *leaf*, preferably with some indication that a leaf is *part* of a tree, rather than for instance a *type* of tree, and iii) it would be at least helpful if it represented facts such as "There are no trees in the Antarctic."

Having said that, there is a large spectrum of complexity and ambition amongst ontology projects. One measure of complexity is the logical expressivity of the relevant ontology language [McGuinness 2003], which has a direct trade-off with inferential tractability, due to the vastly increased computation required to prove statements true in more expressive languages. Expressivity ranges from thesaurus-style representations of synonyms and homonyms, through frame-systems in which individuals are placed in classes in a subsumption hierarchy, through description logics that constitute large decidable fragments of first-order logic [Baader *et al*. 2007], to full first-order and even higher-order logic—for instance the Cyc project, with its purpose-built inference engine [Lenat 1995]. Ontology work began in earnest in the 1980s as a branch of AI research. After an initial rush of enthusiasm, the trade-off between logical expressivity and inferential tractability emerged and became a major obstacle, because much of the human knowledge that arguably should be represented in an ontology can only be stated in languages of great logical expressivity—for instance, negations and disjunctions require full first-order logic, while statements about statements require higher-order logic. Nevertheless, the goals of formal ontology have reawakened with the semantic web [Berners-Lee *et al*. 2001; Berners-Lee 2003]. Since Berners-Lee's vision is to index the web via *meanings,* not just character-strings, it is widely accepted that it will have to draw on some kind of shared, machine-readable, conceptual scheme. But the big stumbling block has been obtaining the world's involvement. At least two major problems need to be solved—first to define "semantic metadata" and then to mark up the web with it.

The World-Wide Web Consortium recently defined a web ontology language, OWL [McGuinness and van Harmelen 2004]. It has three versions of different levels of expressivity: Owl Lite (thesaurus level), OWL DL (description logic-level) and OWL Full (full first-order logic). But attempts to set up repositories for large-scale sharing and

re-use of OWL ontologies have failed to gain traction. It is worth emphasizing that the manual creation of ontologies is enormously difficult. It requires detailed knowledge of formal logic, and for the creation of upper and middle ontologies some understanding of metaphysics (whether explicitly formulated or "quick and dirty"). Moreover, as size increases, so do the interconnections amongst ontology's categories, rendering the potential ramifications of local changes exponentially more significant. Cyc, the most ambitious ontology project, has employed specialist ontological engineers with PhDs in philosophy over a period of 20 years without reaching any natural end-point to the development process. Its nearest competitor, SUMO,[27] is an order of magnitude smaller. Large ontologies have been created for specific, well-funded research areas such as biomedical science, e.g. the Gene Ontology[28] and SNOMED,[29] but again with a huge investment of labor. They are not without their problems [Smith *et al.* 2003], and have to be continually updated. Projects in 'ontology learning' have been tried but so far achieved rather poor performance [Buitelaar 2005].

Could Wikipedia, with its abundance of free, up-to-the-minute contributions, high visibility and remarkable consensus, be used to bypass these laborious ontology-creation methods? Section 2.3.5 mentioned ways in which it may already be seen in this light: its articles are basic concepts, both general concepts and named entities, arranged in some kind of hierarchy via the category structure, and further organisable via a wealth of other relations that may be mined from Wikipedia's structure. There is a vast quantity of "domain-ontology" facts in structured and semi-structured form**.** On the downside, however, as noted in Section 2.2.6, Wikipedia's category system seems currently incapable of supporting principled knowledge inheritance, on pain of, for instance, inferring *isA*(*Domestic Pig, Pork*). Finally, Wikipedia provides no means to perform inferences over its various structures.

This section, like Section 5, is organized around the different kinds of features that researchers seek to mine from Wikipedia. However, because the task is now ontology-building, we consider a somewhat different list, namely: *knowledge organization, named entities*, *synonymy relations and other thesaurus-type information*, *ontology alignments* and finally full-blown *facts*. This research area may alternatively be broken down into projects that seek to augment already existing ontologies or knowledge bases, including Wikipedia itself, and those that build brand new resources, and we will see both kinds.

---

[27] *http://www.ontologyportal.org*
[28] *http://www.geneontology.org*
[29] *http://www.snowmed.org*

6.2 Knowledge Organization

Halavais and Lackaff [2008] assess the overall breadth and comprehensiveness of Wikipedia's coverage of all knowledge. They ask whether the particular enthusiasms of volunteer editors produce excessive coverage of certain topics by comparing topic-distribution in Wikipedia with that in *Books In Print*, and with a range of printed scholarly encyclopedias. They measure this using a Library of Congress categorization of 3000 randomly-chosen articles and find Wikipedia's coverage remarkably representative, except for law and medicine.

Muchnik *et al.* [2007] recommend automatic generation of knowledge hierarchies. They develop five algorithms for organizing Wikipedia articles into a hierarchy, which they evaluate against Wikipedia's category hierarchy. They note that although the matches are not exact, the category hierarchy itself leaves much to be desired—it would be fruitful to evaluate both against human benchmarks.

6.3 Named Entities

Turning now to *named entities*, Section 3.2.2 described detailed methods for disambiguating named entity terms by linking them to Wikipedia articles; Section 4.4 covered named entity ranking for question answering; and Section 5.3 looked at ways of recognizing named entities in Wikipedia itself.

Here it is worth highlighting Wikipedia's natural and straightforward role as *indexer* of named entities. Regarding Wikipedia article URLs as URIs solves one of the most significant problems facing the semantic web: it is easy to create a XML/RDF namespace that names an entity, but difficult to publicize this URI, get anyone else to use it, or coordinate with other possible definitions of namespaces to represent the same things [Legg 2007]. Many authors have noted that Wikipedia, by contrast, enjoys all the broad acceptance and availability that semantic web proponents originally hoped for (e.g. Hepp *et al.* [2006], Bhole *et al.* [2007], McCool [2006]). However, using named entity URIs for semantic web purposes arguably awaits the arrival of URIs for further crucial features of human language, such as general terms (e.g. *tree*), and predicates (e.g. *cut down*).

6.4 Thesaurus Information

Section 3 discussed mining Wikipedia for 'thesaurus-style information'—namely semantic relatedness measures (Section 3.1) and word sense disambiguation (3.2). Here we specifically discuss the use of Wikipedia to generate large-scale, independent, general and systematic thesauri. There is a natural bridge from this task to full-blown ontology-building, for once a system of terms is interconnected via links representing general

semantic relatedness, these links may then be upgraded, or 'typed', to more specific ontological relations.

Gregorowicz and Kramer [2006] seek to construct a comprehensive term-concept map that will solve "the problem of variable terminology" and facilitate concept-based information retrieval by resolving synonyms in a systematic way. They use all Wikipedia articles as concepts, and establish synonyms via redirects and homonyms via disambiguation pages. The result is 2M concepts linked to 3M terms—a vast and impressive resource compared to WordNet's 115,000 synsets created from 150,000 words. Likewise Nakayama *et al.* [2007, 2007, 2008] describe a project to build a large general-purpose thesaurus solely from Wikipedia's hyperlink structure, obtaining a thesaurus of 1.3M concepts with a measured strength of relatedness between each one. They then suggest upgrading the thesaurus to a full-blown ontology by typing the generic relatedness measures between concepts into more traditional ontological relations such as *isA* and *partOf*. Details of how this will be done are sketchy.

The idea of link typing is developed in greater detail in [Krötzsch *et al.* 2005, 2007] and [Völkel *et al.* 2006]. Unlike Nakayama *et al.*, however, they plan to apply it to Wikipedia's own hyperlink structure. They note the profusion of links between articles, all indicating some form of semantic relatedness, and then claim that categorizing them would be a simple, unintrusive way of rendering large parts of Wikipedia machine-readable. For instance, the existing hyperlink from *Leaf* to *Plant* would be labeled *partOf*, that from *Leaf* to *Organ* labeled *kindOf*, and so on. Categorizing all hyperlinks would be a significant task, and they recommend introducing a system of link types and encouraging the Wikipedia editors to start using them, and to suggest further types.

This raises interesting usability issues. Given that ontology is specialist knowledge (at least as traditionally practiced by ontological engineers), it might be argued that disaster could result if every Wikipedian were allowed to apply it in accord with Wikipedia's uniquely democratic editing model. On the other hand, one might ask why this is any different to other specialist additions to Wikipedia (e.g. cell biology, diesel locomotive engineering, Scottish jazz musicians), whose contributors show a remarkable ability to self-select, yielding surprising and impressive quality control. Perhaps the most tricky characteristic of ontology is that, unlike specialist topics such as cell biology, people think they are experts in it when in fact they are not. At any rate, this research is essentially a proposal for Wikipedia's developers to add further functionality, and its results cannot yet be evaluated.

Like Krötzsch *et al.*, Wu and Weld [2007, 2008] seek to augment Wikipedia itself. Their aim is to help kick-start the semantic web by marking up Wikipedia semantically

in order to create enough structured data to make it worthwhile for developers to produce applications for it. To do this they propose a combination of automated and human processes. They investigate the use of machine learning techniques for completing infoboxes by extracting data from article text, constructing new infoboxes from templates where appropriate, rationalizing tags, merging replicated data using microformats, disambiguating links, adding additional links, and flagging items for verification, correction, or the addition of missing information. As with Krötzsch *et al.*, it will be interesting to see whether Wikipedia editors will be eager to work on the collaborative side of this project, and also how effective they are. Furthermore, it is worth asking—even if these projects' aims were achieved and Wikipedia became a complete machine-readable knowledge base, would this bring about the semantic web? How exactly would its existence render the rest of the web machine-readable?

Publications from EMLR that were discussed in detail in Section 5.2 may also be viewed under this heading of link-typing for ontology-building. We saw that these authors focused initially on Wikipedia's category network, aiming to discriminate between *isA* and *notIsA* links [Ponzetto and Strube 2007]. They then further discriminated between two kinds of *isA*: class instance and subclass relationships [Zirn *et al*. 2008]. Unlike Krötzsch *et al.*, and Wu and Weld, they seek to accomplish this task entirely automatically by deducing such relations from an analysis of the titles of interlinked categories. How do their results measure up as an ontology? They claim to derive 105,000 *isA* links, roughly one for each Wikipedia category. Evaluation of Zirn *et al*'s results against the entirely manually created ResearchCyc yielded an accuracy of around 83%, which is impressive. However, though large and comparable with Cyc, this is still much smaller than the 2M concepts in Wikipedia's articles. Also, as a mere *isA* taxonomy it constitutes a relatively inexpressive frame-system-level ontology, lacking in any further relations that might define the concepts in the hierarchy. Finally, though it has been released as a giant set of RDF triples, no ready means to perform inferencing over it seems yet available.

Section 5.2 also described how the same research group turned in later work to parsing category titles and using them to derive new (typed) relations between Wikipedia articles [Nastase and Strube 2008]. Because this work qualifies as mining 'facts' for ontology-building purposes, it is discussed in Section 6.6.

6.5 Ontology Alignment

Finding categories in different ontologies that in some sense "mean the same" can be a useful exercise in itself. If the resources are in the same language, string-matching on category titles goes a long way but is insufficient: homonyms in the mappings must be

detected and eliminated. This task thus overlaps greatly with the word sense disambiguation problem discussed in Section 3.2. The problem cuts both ways: there may be one-to-many string matches from a concept in either of the mapped ontologies to concepts in the other.

WordNet is a popular choice of ontology for alignment projects because it is simple and fairly large (frame-system level). Thus, as was described in Section 3.2.3, Ruiz-Casado *et al.* [2005] align Wikipedia articles with WordNet synsets, building a large general resource that marks up synsets with article URIs and bags of words from article text. However, other than the mapping itself this project adds no ontological value to WordNet, particularly since Wikipedia entries whose title string does not already appear in a synset were discarded. The authors' later work (described in Section 5.1) has shifted to extracting semantic relationships. Suchanek *et al.* [2007, forthcoming] also align WordNet and Wikipedia. However, discussion is deferred to Section 6.6 because they add many other relations as well.

Medelyan and Legg [2008] map 50,000 Wikipedia articles to equivalent categories in ResearchCyc. Their ultimate aim is to create a resource combining Cyc's principled ontological structure with Wikipedia's messier but much more abundant information. Instead of selecting one resource as a base, they merely produce a list of pairs of equivalent concepts in both resources. They use methods described in Section 3.2.3 to determine genuine semantic similarity, following earlier work aligning a domain-specific thesaurus (Agrovoc) with Wikipedia [Medelyan and Milne 2008]. For each Cyc term, its surrounding ontology is used to gather a context for disambiguation, using the taxonomic relations *#$genls*, *#$isa* and some specific relations like #$*countryOfCity* and *#$conceptuallyRelated*. Then the most common Wikipedia article for each context term is identified and compared with all candidates for a mapping. A further test is applied when several Cyc terms map to the same Wikipedia article—reverse disambiguation. First, mappings that score less than 30% of the highest score are eliminated. Then a common-sense test is applied to the remainder based on Cyc's ontological knowledge regarding disjointness between classes. If the best scoring Cyc term does not intersect with the second best one (that is, it represents "a different kind of thing"), the latter is eliminated; otherwise both mappings are accepted. An evaluation on 10,000 manually mapped terms provided by the Cyc Foundation, as well as a study with six human subjects, shows that performance of the mapping algorithm compares with the efforts of humans.

Relation	Domain	Range	Number of facts
subClassOf	class	class	143,210
type	entity	class	1,901,130
context	entity	entity	40,000,000
describes	word	entity	986,628
bornInYear	person	year	188,128
diedInYear	person	year	92,607
establishedIn	entity	year	13,619
locatedIn	object	region	59,716
writtenInYear	book	year	9,670
politicianOf	organization	person	3,599
hasWonPrize	person	prize	1,016
means	word	entity	1,598,684
familyNameOf	word	person	223,194
givenNameOf	word	person	217,132

Table 5. Size of YAGO (facts).

6.6 Facts

Now we turn to mining Wikipedia for what might be called full-blown *facts*, for the purpose of ontology building. This category is blurred by the difficulty of defining what exactly constitutes a fact—e.g., the typing of links in Section 6.4 in some sense already qualifies. However, here we focus on projects that find and store entirely new literals, RDF triples and similar propositionally-structured entities. Sections 4 and 5 have covered much of this work; here we consider to what extent it has resulted in large-scale re-usable knowledge resources.

First we consider those who use Wikipedia to add facts to existing ontologies. We saw in Section 5.2 that Suchanek *et al.* [2007; forthcoming] use information extraction methods to create an ontology named YAGO[30] that unifies WordNet and Wikipedia. This contains 1M concepts and 5M facts about them, an impressive quantity. Table 5 breaks down the number of different types of fact. The concepts are all WordNet synsets, Wikipedia leaf categories and all Wikipedia articles whose titles are not listed as common names in WordNet. This neatly bypasses the poor ontological quality of Wikipedia's category structure, WordNet's taxonomy being manually generated and far cleaner. It also avoids Ruiz-Casado *et al.*'s problem of omitting Wikipedia concepts whose titles do not appear in WordNet, although it still misses all proper names with WordNet *synonyms*—e.g. the programming language Python and the movie *The Birds*. In this way a graph-structured hierarchy of concepts is established, then embellished with facts harvested by a sophisticated suite of heuristics, many obtained by hand-picking popular patterns in the titles of Wikipedia categories and assigning relevant facts to all the instances of those categories. From an ontology-building perspective, these

[30] *http://www.mpi-inf.mpg.de/~suchanek/downloads/yago/*

Dataset	Description	Triples
Page links	Internal links between DBpedia instances derived from the internal pagelinks between Wikipedia articles	62 M
Infoboxes	Data attributes for concepts that have been extracted from Wikipedia infoboxes	15.5 M
Articles	Descriptions of all 1.95 million concepts within the English Wikipedia. Includes titles, short abstracts, thumbnails and links to the corresponding articles	7.6 M
Languages	Additional titles, short abstracts and Wikipedia article links in 13 other languages.	5.7 M
Article categories	Links from concepts to categories using SKOS	5.2 M
Extended abstracts	Additional, extended English abstracts	2.1 M
Language abstracts	Extended abstracts in 13 languages	1.9 M
Type information	Inferred from category structure and redirects by the YAGO ("yet another great ontology") project [Suchanek et al. 2007]	1.9 M
External links	Links to external web pages about a concept	1.6 M
Categories	Information which concept is a category and how categories are related	1 M
Persons	Information about 80,000 persons (date and place of birth etc.) represented using the FOAF vocabulary	0.5 M
External links	Links between DBpedia and Geonames, US Census, Musicbrainz, Project Gutenberg, the DBLP bibliography and the RDF Book Mashup	180 K

Table 6. Content of DBPedia [Auer et al. 2007].

sophisticated automated methods are a real step forward, though only a tiny subset of category names has been parsed. For instance they do not address widespread patterns such as "X by Y" (e.g. *Persons by continent*, *Persons by company*, *Persons by nationality* and so on), which was analyzed by the EMLR group (Section 5.2).

YAGO has many features one seeks in a formal ontology. Its authors have defined a logic-based representation language and a basic data model of entities and binary relations, with a small extension to represent relations between facts (such as transitivity). This gives it formal rigor—the authors even provide a model-theoretic semantics—and the expressive power of a rich version of Description Logic. In terms of inferential tractability it compares favorably with the hand-crafted Cyc. A SPARQL interface (available online) allows queries of traditional knowledge-base logical complexity—for instance when asked for billionaires born in the USA it came up with two (though it missed *Bill Gates*—coverage of Wikipedia's structured data is not complete by the project's methods). The authors plan to integrate their project with the latest version of OWL (released in 2007). They claim to have already noticed a positive feedback loop whereby as more facts are added, word senses can be disambiguated more effectively in order to correctly identify and enter further facts. Such a feedback loop was a long-standing ambition of AI researchers (e.g. Lenat [1995]), though claims that it was about to be achieved often turned out to be premature.

By contrast, the flourishing and ambitious DBpedia project [Auer *et al.* 2007; Auer and Lehmann 2007] attempts to create an entirely new ontology by harvesting facts from Wikipedia. The facts are stored as a vast set of RDF triples. As noted in Section 5.2, this project strives to make all Wikipedia's structured information freely available in database form. Of all projects, it takes the most purely automated approach and gathers the largest quantity of structured data. The focus is on formatting patterns in the text of Wikipedia articles, notably infoboxes, though categorization and other links are also harvested. A staggering 103M "facts" (triplets) are obtained. Like YAGO, the dataset can be queried via SPARQL and Linked Data, and connects with other open datasets on the web. Table 6 summarizes its content.

The project has already been influential—for instance, to test their document classification algorithm Janik and Kochut [2007] use slightly modified methods from DBpedia to create an RDF ontology from Wikipedia (Section 4.5). From a general ontology-building perspective, however, it has some weaknesses. There is little or no connection between the facts, and the knowledge is not organized into a hierarchy that enables inheritance (although, of course, as a giant database, state of the art processing techniques can be brought to bear). Unlike YAGO it has no formally defined ontology language, and thus it would seem that many semantic relations amongst its triples will go unrecognized (e.g., that the first argument of the predicate *artistOf* might bear a relationship to the collection *Artists*). Second, although a formal evaluation of the resource's quality is not provided, a quick manual inspection reveals that large sections of the data has limited ontological value. For instance, 60% of the RDF triples are internal links derived from Wikipedia's link structure; only 15% are taken directly from infoboxes, and of those, the most common relation (over 10%) is the formatting relation *wikiPageUsesTemplate*. Amongst the properly ontological relations are many obvious redundancies not identified as such, e.g. *placeOfBirth* and *birthPlace*, *dateOfBirth* and *birthDate*. Finally, some individual relations contain poor-quality infobox data—for instance, *keyPeople* assertions of the form "CEO" or "Bob".

We finally come to consider the final phase of EMLR's project [Nastase and Strube 2008]. We saw in Section 5.2 that this work consisted in parsing category titles, analyzing patterns in them and using that information to derive new relations between articles. They manage a deeper analysis of category titles than YAGO—in particular, they managing to crack open the extensive *X by Y* pattern and derive entirely implicit relations, as we saw above. In this way they manage to add a wealth of new ontological information to their existing taxonomy of 105,000 categories—9M new facts, about twice the size of YAGO. The facts include 3.4 million *isA* and 3.2 million *spatial* relations,

	Ontology	Entities	Facts
Manually created	SUMO	20,000	60,000
	WordNet	117,597	207,016
	OpenCyc	47,000	306,000
	ResearchCyc	250,000	2,200,000
Automatically derived	YAGO	1M	5M
	DBpedia	N/A	103M
	EMLR[2008]	105,000	9M

Table 7. Size of ontologies (adapted from Suchanek *et al.* [2007]).

along with 43,000 *memberOf* relations and 44,000 other specific relations such as *causedBy* and *writtenBy*. The authors promise to release a new ontology containing these facts soon. It will be interesting to see whether they define a formally specified ontology language, as with YAGO (and if so how expressive it is), or merely dump out the data as with DBpedia (in which case the tools available for inferencing, and the complexity of supported queries, become paramount).

Table 7 shows the size of the larger ontologies. How much nearer does this work bring us to the semantic web? Great progress has been made on named entities (such as 'Helen Clark'), for all that is needed to establish shared meaning for a named entity is a shared URI. General concepts (such as 'tree') are more tricky. There is certainly a wealth of semantic information regarding such concepts in Wikipedia, but an almost total lack of consensus on how to extract and analyze it, let alone inference over it. Yet for the semantic web, this was the whole point.

7. PEOPLE, PLACES AND RESOURCES

The research described here is scattered across the globe; Figure 16 shows prominent countries and institutions.

US and Germany are the largest contributors. The US research spreads across many institutions. The University of North Texas, who work with entity recognition and disambiguation, produced the *wikify* system. In the Pacific Northwest, Microsoft Research focuses on named entity recognition, while the University of Washington extracts semantic relations from Wikipedia's infoboxes. German research is more localized geographically. EML Research Institute works on relation extraction, semantic relatedness, and co-reference resolution; Darmstadt University of Technology on semantic relatedness and analyzing Wikipedia's structure. The Max-Plank Institut produced the YAGO ontology; they collaborate with the University of Leipzig, who produced DBpedia. The University of Karlsruhe have focused on providing users with tools to add formal semantics to Wikipedia.

Spain is Europe's next largest contributor. Universidad Autonoma de Madrid extract semantic relations from Wikipedia; Universidad Politecnica de Valencia and Universidad de Alicente both use it to answer questions and recognize named entities. The Netherlands, France, and UK are each represented by a single institution. The University of Amsterdam focusses on question answering; INRIA works primarily on entity ranking, and Imperial College on recognizing and disambiguating geographical locations.

The Israel Institute of Technology have produced widely cited work on semantic relatedness, document representation and categorization. They developed the popular technique of Explicit Semantic Analysis.

Hewlett Packard's branch in Bangalore puts India on the map with document categorization research. In China, Shanghai Jiatong University works on relation extraction and category recommendation. In Japan, the University of Osaka has produced several open source resources, including a thesaurus and a bilingual (Japanese–English) dictionary. The University of Tokyo, in conjunction with the National Institute of Advanced Industrial Science and Technology, have focused on relation extraction.

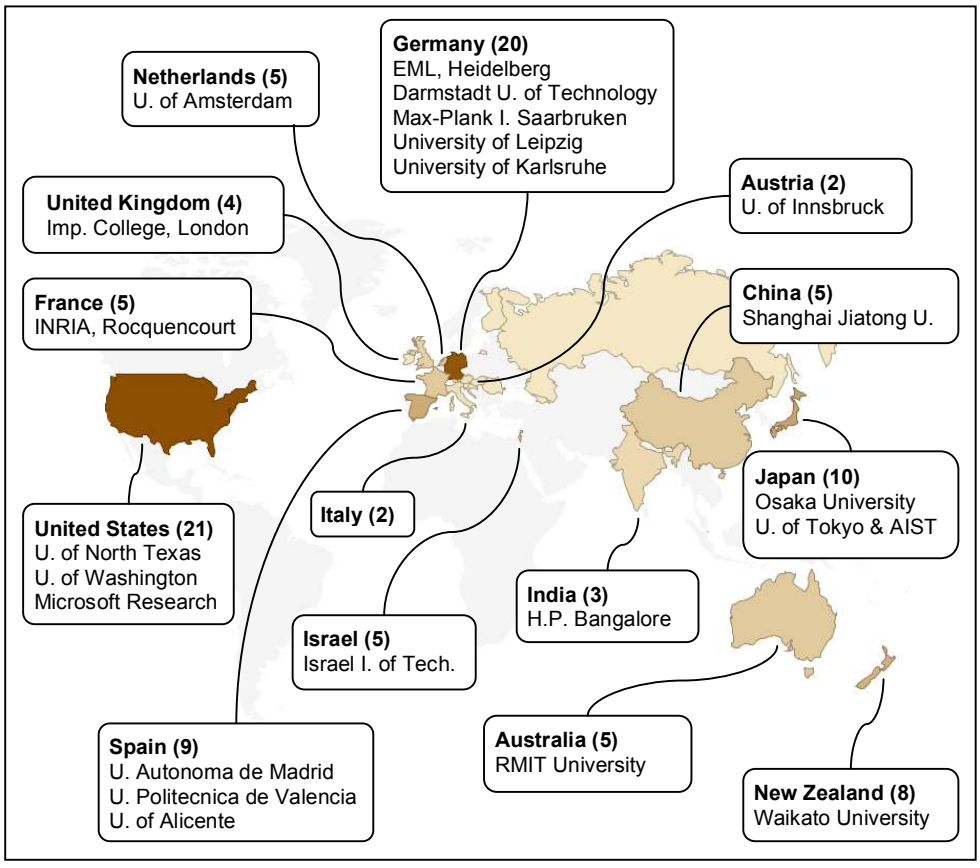

Figure 16. Countries and institutions with significant research on mining meaning from Wikipedia.

New Zealand and Australia are each represented by a single institution. Research at the University of Waikato covers entity recognition, query expansion, topic indexing, semantic relatedness and augmenting existing knowledge bases. RMIT in Melbourne have collaborated with INRIA's work on entity ranking.

Table 8 summarizes tools and resources, along with brief descriptions and URLs. The first part shows tools for accessing and processing Wikipedia. The second shows demos of Wikipedia mining applications. The third lists datasets that have been generated from Wikipedia.

Processing tools	
JWPL Java Wikipedia Library	API for structural access of Wikipedia parts such as redirects, categories, articles and link structure. [Zesch et al. 2008] *http://www.ukp.tu-darmstadt.de/software/jwpl/*
WikiRelate!	API for computing semantic relatedness using Wikipedia [Strube and Ponzetto 2006; Ponzetto and Strube 2006] *http://www.eml-research.de/ english/research/ nlp/download/ wikipediasimilarity.php*
Wikipedia Miner	API that provides a simplified access to Wikipedia and models its structure semantically [Milne et al. 2008] *http://sourceforge.net/ projects/wikipedia-miner/*
WikiPrep	A Perl tool for preprocessing Wikipedia XML dumps [Gabrilovich and Markovitch 2007] *http://www.cs.technion.ac.il/ ~gabr/resources/ code/wikiprep/*
W.H.A.T. Wikipedia Hybrid Analysis Tool	An analytic tool for Wikipedia with two main functionalities: an article network and extensive statistics. It contains a visualization of the article networks and a powerful interface to analyze the behavior of authors. *http://sourceforge.net/ projects/ w-h-a-t/*

Wikipedia mining demos	
DBpedia Online Access	Online access of DBpedia data (103M facts extracted from Wikipedia) via a SPARQL query endpoint and as Linked Data. [Auer et al. 2007] *http://wiki.dbpedia.org/ OnlineAccess*
YAGO	Demo of the Yet Another Ontology YAGO, containing 1.7M entities and 14M facts [Suchanek et al. 2007] *http://www.mpii.mpg.de/ ~suchanek/yago*
QuALiM	A Question Answering system. Given a question in a natural language returns relevant passages from Wikipedia. [Kaisser 2008] *http://demos.inf.ed.ac.uk:8080/ qualim/*
Koru	A demo of a search interface that maps topics involved in both queries and documents to Wikipedia articles. Supports automatic and interactive query expansion. [Milne et al. 2007] *http://www.nzdl.org/koru*
Wikipedia Thesaurus	A large scale association thesaurus containing 78 million associations [Nakayama et al. 2007 and 2008] *http://wikipedia-lab.org:8080/ WikipediaThesaurusV2/*
Wikipedia English-Japanese	A dictionary returning translations from English into Japanese and vise versa, enriched with probabilities of these translations [Erdmann et al. 2007]

dictionary	*http://wikipedia-lab.org:8080/ WikipediaBilingualDictionary/*
Wikify	Automatically annotates any text with links to Wikipedia articles [Mihalcea and Csomai 2007]
	http://wikifyer.com/
Wikifier	Automatically annotates any text with links to Wikipedia articles describing named entities
	http://wikifier.labs.exalead.com/
Location query server	Location data accessible via REST requests returning data in a SOAP envelope. Two requests are supported: A bounding box or a Wikipedia Article. The reply is the number of references made to locations within that bounding box, and a list of Wikipedia articles describing those locations. Or none, if the request is not a location. [Overell and Rüger 2006 and 2007]
	http://www.doc.ic.ac.uk/ ~seo01/wiki/demos
Datasets	
DBpedia	Facts extracted from Wikipedia infoboxes and link structure in RDF format. [Auer et al. 2007]
	http://wiki.dbpedia.org
Wikipedia Taxonomy	Taxonomy automatically generated from the network of categories in Wikipedia (RDF Schema format) [Ponzetto and Strube 2007; Zirn et al. 2008]
	http://www.eml-research.de/ english/research/ nlp/download/ wikitaxonomy.php
Semantic Wikipedia	A snapshot of Wikipedia automatically annotated with named entity tags. [Zaragossa et al. 2007]
	http://www.yr-bcn.es/ semanticWikipedia
Cyc to Wikipedia mappings	50,000 automatically created mappings from Cyc terms to Wikipedia articles. [Medelyan and Legg 2008]
	http://www.cs.waikato.ac.nz/ ~olena/cyc.html
Topic indexed documents	A set of 20 Computer Science technical reports indexed with Wikipedia articles as topics. 15 teams of 2 senior CS undergraduates have independently assigned topics from Wikipedia to each article. [Medelyan et al. 2008]
	http://www.cs.waikato.ac.nz/ ~olena/wikipedia.html
Locations in Wikipedia, ground truth	A manually annotated sample of 1000 Wikipedia articles. Each link in each article is annotated, whether it is a location or not. If yes, it contains the corresponding unique id from the TGN gazetteer. [Overell and Rüger 2006 and 2007]
	http://www.doc.ic.ac.uk/ ~seo01/wiki/data_release

Table 8. Wikipedia tools and resources.

8. SUMMARY

A whole host of researchers have been quick to grasp the potential of Wikipedia as a resource for mining meaning: the literature is large and growing rapidly.

We began this article by describing Wikipedia's creation process and structure (Section 2). The unique open editing philosophy, which accounts for its success, is subversive. Although regarded as suspect by the academic establishment, it is a remarkable concrete realization of the American pragmatist philosopher Peirce's proposal that knowledge be defined through its public character and future usefulness rather than

any prior justification. Wikipedia is not just an encyclopedia but can be viewed as anything from a corpus, taxonomy, thesaurus, hierarchy of knowledge topics to a full-blown ontology. It includes explicit information about synonyms (redirects) and word senses (disambiguation pages), database-style information (infoboxes), semantic network information (hyperlinks), category information (category structure), discussion pages, and the full edit history of every article. Each of these sources of information can be mined in various ways.

Section 3 explains how Wikipedia is being drawn upon for natural language processing. Unlike WordNet, it was not created as a lexical resource that reflects the intricacies of human language. Instead, its primary goal is to provide encyclopedic knowledge across subjects and languages. However, the research described here demonstrates that it has, unexpectedly, immense potential as a repository of linguistic knowledge for natural language applications. In particular, its unique features allow well-defined tasks such as word sense disambiguation and word similarity to be addressed automatically—and the resulting level of performance is remarkably high. Researchers on co-reference resolution and mining of multilingual information have only recently discovered Wikipedia; significant improvements in these areas can be expected shortly. To our knowledge, its use as a resource for other tasks such as natural language generation, machine translation and discourse analysis, has not yet been explored. These areas are ripe for exploitation, and exciting discoveries can be expected.

Section 4 describes applications to information retrieval. Query expansion, document classification and topic indexing provide the best examples of applying Wikipedia for searching and organizing document collections. These areas can take advantage of its unique properties while grounding themselves in—and building upon—existing research. In particular, document classification has gathered momentum and significant advances are obtained over the state of the art. Question answering and entity ranking are less well addressed, because they do not seem to take full advantage of Wikipedia: with a few exceptions they simply treat it as just another corpus and thus differ little from previous work. We found little evidence of cross-pollination between this work and the information extraction efforts described in Section 5. Given how closely question answering and entity ranking depend on the extraction of facts and entities, we expect this to become a fruitful line of enquiry.

In Section 5 we turn to information extraction; mining text for topics, relations and facts. Unlike the tasks in Sections 3 and 4, information extraction is not easy to define. Different researchers focus on different kinds of information: we have reviewed research on extracting information about movie directors and soccer players, composers, corporate

descriptions and hierarchical and ontological relations. Techniques range from those developed for standard text corpora to ones that utilize properties such as hyperlinks and category structure. The extracted resources range in size from several hundred to several million relations, but the lack of a common basis for evaluation prevents us from drawing any conclusion as to which approach performs best.

Section 6 discusses the use of Wikipedia for ontology-building. Wikipedia's vast quantity of structured information provides low-hanging fruit for automating this process. Article names can serve as URIs for named entities; hyperlinks and redirects can be mined for large-scale thesauri; the category structure can be treated as encoding taxonomic information (though not always very well); and infoboxes are a rich source of domain knowledge. From the perspective of large-scale general ontology building, the two most impressive projects are YAGO and DBPedia. Which will turn out to be more useful, the large but messy and low-quality DBPedia, or the smaller but more rigorous and accurate YAGO? Meanwhile, EMLR's latest efforts (not yet released) promise to combine some of the greater rigor of the former with the greater size of the latter. We believe that an extrinsic evaluation would be most meaningful, and hope to see these systems compete on a well-defined task in an independent evaluation. It will also be interesting to see to what extent these resources are exploited by other research communities in the future.

Some authors have suggested using Wikipedia editors themselves to perform ontology-building, an enterprise that might be thought of as mining Wikipedia's *people* rather than its *data*. Perhaps they grasp the implications of the underlying driving force behind this massively successful resource better than the rest of us! Only time will tell whether the community is amenable to following such suggestions. The idea of moving to a more structured and ontologically principled Wikipedia raises an interesting question: how will it interact with the public, amateur-editor model? Does this signal the long-awaited emergence of the semantic web? We suspect that, like the success of Wikipedia itself, the result will be something new, something that experts have not foreseen and may not condone. That is the glory of Wikipedia.

ACKNOWLEDGEMENTS

We warmly thank Evgeniy Gabrilovich, Rada Mihalcea, Dan Weld, Fabian Suchanek and the YAGO team for their valuable comments on a draft of this paper. Medelyan is supported by a scholarship from Google, Milne by a New Zealand Tertiary Education Commission Top Achiever Scholarship.

References

ADAFRE, S.F., JIJKOUN, V., AND M. DE RIJKE. [2007] Fact Discovery in Wikipedia. In Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence.

ADAFRE, S.F., AND M. DE RIJKE. [2006] Finding Similar Sentences across Multiple Languages in Wikipedia. In Proceedings of the EACL 2006 Workshop on New Text–Wikis and Blogs and Other Dynamic Text Sources.

ADAFRE, S.F., AND M. DE RIJKE. [2005] Discovering Missing Links in Wikipedia. In Proceedings of the LinkKDD 2005, August 21, 2005, Chicago, IL.

AGROVOC [1995] Multilingual agricultural thesaurus. Food and Agricultural Organization of the United Nations. http://www.fao.org/agrovoc/

AHN, D., JIJKOUN, V., MISHNE, G., MÜLLER, K., DE RIJKE, M., AND S. SCHLOBACH. [2004] Using Wikipedia at the TREC QA Track. In Proceedings of the 13th Text Retrieval Conference (TREC 2004).

ALLAN, J. [2005] HARD track overview in TREC 2005: High accuracy retrieval from documents. In Proceedings of the 14th Text Retrieval Conference (TREC 2005).

AUER, S., BIZER, C., LEHMANN, J., KOBILAROV, G., CYGANIAK, R., AND Z. IVES [2007] DBpedia: A Nucleus for a Web of Open Data. In Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea, 4825: 715–728, 2007.

AUER, S. AND J. LEHMANN. [2007] What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. In Franconi *et al.* (eds), Proceedings of European Semantic Web Conference (ESWC'07), LNCS 4519, pp. 503–517, Springer, 2007.

BAADER, F., CALVANESE, D., MCGUINESS, D. AND D. NARDI. [2007] The Description Logic Handbook: Theory, Implementation and Applications. Cambridge: Cambridge University Press.

BAKER, L. [2008] Professor Bans Google & Wikipedia: Encourages Critical Thinking & Research. Search Engine Journal, January 14th, 2008.

BANERJEE, S. [2007] Boosting Inductive Transfer for Text Classification Using Wikipedia. In Proceedings of the 6th International Conference on Machine Learning and Applications (ICMLA), pp. 148–153.

BANERJEE, S., RAMANATHAN, K. AND A. GUPTA. [2007] Clustering Short Texts using Wikipedia. In Proceedings of the 30th Annual International ACM SIGIR conference on Research and Development in Information Retrieval. Amsterdam, Netherlands. pp. 787–788.

BANKO, M., CAFARELLA, M. J., SODERLAND, S., BROADHEAD, M. AND O. ETZIONI. [2007] Open information extraction from the Web. In Proceedings of the 20th International Joint Conference on Artificial Intelligence IJCAI'07, pp. 2670–2676, January 2007.

BHOLE, A., FORTUNA, B., GROBELNIK, B. AND D. MLADENIĆ. [2007] Extracting Named Entities and Relating Them over Time Based on Wikipedia. Informatica.

BELLOMI, F. AND R. BONATO. [2005] Network Analysis for Wikipedia. In Proceedings of the 1st International Wikimedia Conference, Wikimania 2005. Wikimedia Foundation.

BERNERS-LEE, T., HENDLER, J, AND O. LASSILA. [2001]. The Semantic Web. Scientific American 284 (5), 34–43.

BERNERS-LEE, T. [2003]. Foreword. In D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster (Eds.) Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential. Cambridge, MA: MIT Press.

BRIN, S. AND L. PAGE. [1998] The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, Vol. 33, pp. 107–117.

BROWN, P., DELLA PIETRA, S., DELLA PIETRA, V., AND R. MERCER. [1993] The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19(2), 263–311.

BUDANITSKY, A. AND HIRST, G. [2001] Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, PA.

BUITELAAR, P., CIMIANO, P., MAGNINI, B. (eds). [2005] Ontology Learning from Text: Methods, Evaluation and Applications. Amsterdam, The Netherlands: IOS Press.

BUNESCU, B. AND PAŞCA, M. [2006] Using Encyclopedic Knowledge for Named Entity Disambiguation. In Proceedings of the11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 9–16.

BUSCALDI, D. AND P. A. ROSSO. [2007] Comparison of Methods for the Automatic Identification of Locations in Wikipedia. In Proceedings of the 4th ACM workshop on Geographical information retrieval, GIR'07. Lisbon, Portugal, pp. 89–92.

BUSCALDI, D. AND P. A. ROSSO. [2007] A Bag-of-Words Based Ranking Method for the Wikipedia Question Answering. Task Evaluation of Multilingual and Multi-modal Information Retrieval, pp. 550–553.

CAVNAR, W. B. AND J. M. TRENKLE. [1994] N-Gram-Based Text Categorization. In Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175.

CHERNOV, S., IOFCIU, T., NEJDL, W. AND X. ZHOU. [2006] Extracting Semantic Relationships between Wikipedia Categories. In Proceedings of the 1st International Workshop: SemWiki'06—From Wiki to Semantics. Co-located with the 3rd Annual European Semantic Web Conference ESWC'06 in Budva, Montenegro, June 12, 2006.

CIMIANO, P. AND J. VOLKER. [2005] Towards large-scale, open-domain and ontology-based named entity classification. In Proceedings of the Internatioal Conference on Recent Advances in Natural Language Processing, RANLP'05, pp. 166–172. INCOMA Ltd., Borovets, Bulgaria, September 2005.

CSOMAI, A. AND R. MIHALCEA. [2007] Linking Educational Materials to Encyclopedic Knowledge. Frontiers in Artificial Intelligence and Applications, v.158, pp. 557–559. IOS Press, Netherlands.

CUCERZAN, S. [2007] Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716, Prague, Czech Republic, June 2007.

CULOTTA, A., MCCALLUM, A. AND J. BETZ. [2006]. Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. New York, NY, pp. 296–303.

DAKKA, W. AND S. CUCERZAN. [2008]. Augmenting Wikipedia with Named Entity Tags. In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad.

DENNING, P., HORNING, J., PARNAS, D., AND WEINSTEIN, L. [2005]. Wikipedia Risks. In *Communications of the ACM 48*(12), pp. 152–152.

DENOYER, L. AND GALLINARI, P. [2006] The Wikipedia XML corpus. SIGIR Forum, 40(1), pp. 64–69, ACM Press.

DONDIO, P., BARRETT, S., WEBER, S., AND SEIGNEUR, J. [2006] Extracting Trust from Domain Analysis: A Case Study on the Wikipedia Project. Autonomous and Trusted Computing, pp. 362-373.

DUMAIS, S., PLATT, J., HECKERMAN, D. AND M. SAHAMI. [1998] Inductive learning algorithms and representations for text categorization. In Proceedings of the 7th international conference on Information and knowledge management, pp. 148–155.

EDMONDS, P. AND KILGARRIFF, A. [2002] Introduction to the special issue on evaluating word sense disambiguation systems. Journal of Natural Language Engineering, 8(4), pp. 279–291. Cambridge University Press, New York, NY, USA.

EMIGH, W. AND HERRING, S. [2005] Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias. In Proceedings of the 38th Hawaii International Conference on System Sciences, p.99a.

ERDMANN, M., NAKAYAMA, K., HARA, T., AND S. NISHIO. [2008] An Approach for Extracting Bilingual Terminology from Wikipedia. In Proceedings of the 13th International Conference on Database Systems for Advanced Applications (DASFAA, To appear).

FELLBAUM, C. (editor). [1998] WordNet An Electronic Lexical Database. Cambridge, MA: MIT Press.

FERRÁNDEZ, F., TORAL, A., FERRÁNDEZ, Ó., FERRÁNDEZ, A., AND R. MUÑOZ. [2007] Applying Wikipedia's Multilingual Knowledge to Cross–Lingual Question Answering. In Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems, Paris, France, pp. 352–363. June 2007

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G., AND E. RUPPIN. [2002] Placing search in context: The concept revisited. ACM Transactions on Information Systems, 20(1), pp. 116–131.

FRANK, E., PAYNTER, G. W., WITTEN, I. H., GUTWIN, C. AND C. G. NEVILL-MANNING. [1999] Domain-Specific Keyphrase Extraction. In Proceedings of the 16th International Joint Conference on Artificial Intelligence, IJCAI'99, Stockholm, Sweden, pp. 668–673.

GABRILOVICH, G. AND S. MARKOVITCH. [2007] Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07, Hyderabad, India, January 2007, p.1606–1611.

GABRILOVICH, G. AND MARKOVITCH, S. [2006] Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, Proceedings of The 21st National Conference on Artificial Intelligence (AAAI), pp. 1301–1306, Boston, July 2006

GILES, J. [2005] Internet Encyclopaedias Go Head to Head. In *Nature 138*(15), 14 December 2005.

GLEIM, R., MEHLER, A. AND M. DEHMER. [2007] Web Corpus Mining by Instance of Wikipedia. In Kilgarriff, Adam; Baroni, Marco (eds.) Proceedings of the EACL 2006 Workshop on Web as Corpus, Trento, Italy, April 3–7, 2006, pp. 67–74.

GREGOROWICZ, A. AND M. A. KRAMER. [2006] Mining a Large-Scale Term-Concept Network from Wikipedia. Mitre Technical Report 06–1028, October 2006.

HALAVAIS, A. AND LACKAFF, D. [2008] An Analysis of Topical Coverage of Wikipedia. Journal of Computer-Mediated Communication, 13(2), pp. 429–440.

HALLER, H., KRÖTZSCH, M., VÖLKEL, M., AND D. VRANDECIC. [2006] Semantic Wikipedia (software demo). In Proceedings of the 2006 International Symposium on Wikis, pp. 137–138. ACM Press, August 2006.

HATCHER, E. AND O. GOSPODNETIC. [2004] Lucene in Action. Manning Publications, Greenwich, CT.

HAVELIWALA, T. H. [2003] Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. IEEE transactions on knowledge and data engineering, 15(4), pp. 784–796.

HERBELOT, A. AND A. COPESTAKE. [2006] Acquiring Ontological Relationships from Wikipedia Using RMRS. In Proc. International Semantic Web Conference 2006 Workshop on Web Content Mining with Human Language Technologies, Athens, GA.

HEPP, M., BACHLECHNER, D., AND K. SIORPAES. [2006] Harvesting Wiki Consensus—Using Wikipedia Entries as Ontology Elements. In Proceedings of the 1st International Workshop: SemWiki'06—From Wiki to

Semantics. Co-located with the 3rd Annual European Semantic Web Conference ESWC'06 in Budva, Montenegro, June 12, 2006.

HIGASHINAKA, R., DOHSAKA, K., AND H. ISOZAKI. [2007] Learning to Rank Definitions to Generate Quizzes for Interactive Information Presentation, in Companion Volume to the Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 117–120

HUANG, W.C., TROTMAN, A., AND S. GEVA. [2007] Collaborative Knowledge Management: Evaluation of Automated Link Discovery in the Wikipedia. In Proceedings of the Workshop on Focused Retrieval at SIGIR 2007, July 27, 2007, Amsterdam.

IDE, N. AND J. VÉRONIS (editors). [1998] Word Sense Disambiguation. Special issue of Computational Linguistics, 24(1).

JIANG, J. J. AND D. W. CONRATH, D. W. [1997] Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the 10th International Conference on Research in Computational Linguistics, ROCLING'97. Taiwan.

JIJKOUN, V. AND M. DE RIJKE. [2006] Overview of the WiQA task at CLEF 2006. In: C. Peters *et al.* (editors). Evaluation of Multilingual and Multi-modal Information Retrieval. 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20–22, 2006, Revised Selected Papers, LNCS 4730, pp. 265–274, September 2007

JANIK, M. AND K. KOCHUT. [2007] Wikipedia in Action: Ontological Knowledge in Text Categorization, University of Georgia, Computer Science Department Technical Report no. UGA-CS-TR-07-001.

KAISSER, M. [2008] The QuALiM Question Answering Demo: Supplementing Answers with Paragraphs drawn from Wikipedia. In Proceedings of the ACL-08 HLT Demo Session, Columbus, Ohio, pp. 32–35.

KASNECI, G., SUCHANEK, F.M., IFRIM, G., RAMANATH, M. AND G. WEIKUM. [2007] NAGA: Searching and Ranking Knowledge. In Proceedings of the 24th IEEE International Conference on Data Engineering, ICDE'08, Cancun, Mexico, 7–12 April 2008, pp. 953–962.

KASSNER, L., NASTASE, V., AND M. STRUBE. [2008] Acquiring a Taxonomy from the German Wikipedia. To appear in Proceedings of LREC 2008.

KAZAMA, J. AND K. TORISAWA. [2007] Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 698–707.

KINZLER, D. [2005] WikiSense: Mining the Wiki, v 1.1. In Proceedings of the 1st International Wikimedia Conference, Wikimania 2005. Wikimedia Foundation.

KITTUR, A., SUH., B., PENDLETON, B.A. AND CHI, E.H. [2007] He says, she says: Conflict and Coordination in Wikipedia. In CHI, pp. 453-462.

KLEINBERG, J. [1998] Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46, pp. 604–632.

KLAVANS, J. L. AND P. RESNIK. [1996] The balancing act: combining symbolic and statistical approaches to language. Cambridge, MA: MIT Press.

KRIZHANOVSKY, A. [2006] Synonym Search in Wikipedia: Synarcher. In Proceedings of the 11th International Conference "Speech and Computer" SPECOM'06. Russia, St. Petersburg, June 25–29, 2006, pp. 474–477.

KRÖTZSCH, M., VRANDECIC, D., VÖLKEL, M., HALLER, H., AND R. STUDER. [2007] Semantic Wikipedia. Journal of Web Semantics, 5, pp. 251–261.

KRÖTZSCH, M., VRANDECIC, D. AND M. VÖLKEL. [2005] Wikipedia and the Semantic Web—The Missing Links. In Proceedings of the 1st International Wikimedia Conference, Wikimania 2005. Wikimedia Foundation.

LEACOCK, C., AND M. CHODOROW. [1998] Combining local context and WordNet similarity for word sense identification. In Fellbaum, C. (editor), WordNet: An Electronic Lexical Database. Chapter 11, pp. 265–283. Cambridge, MA: MIT Press.

LEHTONEN, M. AND A. DOUCET. [2007] EXTIRP: Baseline Retrieval from Wikipedia. Comparative Evaluation of XML Information Retrieval Systems, pp. 115–120.

LEGG, C. [2007] Ontologies on the Semantic Web. Annual Review of Information Science and Technology 41, pp. 407–452.

LENAT, D. B. [1995] Cyc: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 38(11).

LIPSCOMB, C.E. [2000] Medical Subject Headings (MeSH). In Bulletin of the Medical Library Association 88(3), p. 265.

LI, B., CHEN, Q., YEUNG, D.S., NG, W.W.Y., WANG, X. [2007] Exploring Wikipedia and Query Logs Ability for Text Feature Representation. In Proceedings of the International Conference on Machine Learning and Cybernetics, Hong Kong, 19–22 August 2007, v. 6, pp. 3343–3348.

LI, Y., LUK, R. W. P., HO, E. K. S., CHUNG, K. F. [2007] Improving weak ad-hoc queries using Wikipedia as external corpus. In Kraaij *et al.* (editors) Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07, Amsterdam, The Netherlands, July 23–27, 2007, pp. 797–798. ACM Press.

LIH, A. [2004] Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Source. In Proceedings of the 5th International Symposium on Online Journalism.

MAGNUS, P. D. [2006] Epistemology and the Wikipedia. In Proceedings of the North American Computing and Philosophy Conference, Troy, New York, August 2006.

MAYS, E., DAMERAU, F. J. AND R. L. MERCER. [1991] Context-based spelling correction. Information Processing and Management 27(5), pp. 517–522.
MCCOOL, R. [2006]. Rethinking the Semantic Web, Part 2. IEEE Internet Computing 10(1), pp. 93–96.
MCGUINNESS, D. [2003]. Ontologies Come of Age. In D. Fensel, *et al.* (editors) Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. Cambridge, MA: MIT Press.
MCGUINNESS, D. AND F. VAN HARMELEN. [2004] OWL Web Ontology Language: Overview. http://www.w3.org/TR/owl-features/
MEDELYAN, O. AND D. MILNE. [2008] Augmenting domain-specific thesauri with knowledge from Wikipedia. In Proceedings of the NZ Computer Science Research Student Conference, Christchurch, NZ.
MEDELYAN, O., WITTEN, I. H., AND D. MILNE. [2008] Topic Indexing with Wikipedia. To appear in Proceedings of the WIKI-AI: Wikipedia and AI Workshop at the AAAI'08 Conference, Chicago, US.
MEDELYAN, O. AND C. LEGG. [2008] Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. To appear in Proceedings of the WIKI-AI: Wikipedia and AI Workshop at the AAAI'08 Conference, Chicago, US.
MIHALCEA, R. [2007] Using Wikipedia for Automatic Word Sense Disambiguation. In Proceedings of the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, New York, April 2007
MIHALCEA, R. AND D. MOLDOVAN. [2001] Automatic generation of a coarse grained WordNet. In Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources. Pittsburgh, PA.
MIHALCEA, R. AND A. CSOMAI. [2007] Wikify! Linking Documents to Encyclopedic Knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM'07, Lisbon, Portugal, November 6–8, 2007, pp. 233–241.
MILLER, E. [1998] An Introduction to the Resource Description Framework. Bulletin of the American Society for Information Science 25(1), pp. 15–19.
MILLER, G. A., AND W. G. CHARLES. [1991] Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1), pp. 1–28.
MILNE, D., MEDELYAN, O. AND I. H. WITTEN. [2006] Mining domain-specific thesauri from Wikipedia: A case study. In Proceedings of the International Conference on Web Intelligence (IEEE/WIC/ACM WI'2006), Hong Kong.
MILNE, D., WITTEN, I. H. AND D. M. NICHOLS. [2007] A Knowledge-Based Search Engine Powered by Wikipedia. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM'07, Lisbon, Portugal, November 6–8, 2007, pp. 445–454.
MILNE, D. [2007] Computing Semantic Relatedness using Wikipedia Link Structure. In Proceedings of the New Zealand Computer Science Research Student Conference, NZ CSRSC'07, Hamilton, New Zealand.
MILNE D. AND I. H. WITTEN. [2008] Learning to link with Wikipedia. Forthcoming
MINIER, Z., ZALAN, B. AND L. CSATO. [2007] Wikipedia-Based Kernels for Text Categorization. In Proceedings of the International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC'07, IEEE Computer Society Washington, DC, USA. pp. 157–164.
MUCHNIK, L., ITZHACK, R., SOLOMON, S. AND Y. LOUZOUN. [2007] Self-emergence of Knowledge Trees: Extraction of the Wikipedia Hierarchies, in Physical Review E 76(1).
NAKAYAMA, K., HARA, T., AND S. NISHIO. [2007] Wikipedia: A New Frontier for AI Researches. Journal of the Japanese Society for Artificial Intelligence 22(5), pp. 693–701.
NAKAYAMA, K., HARA, T., AND S. NISHIO. [2008] A Search Engine for Browsing the Wikipedia Thesaurus. In Proceedings of the 13th International Conference on Database Systems for Advanced Applications, Demo session (DASFAA'08), pp. 690–693.
NAKAYAMA, K., ITO, M., HARA, T. AND S. NISHIO. [2008] Wikipedia Mining for Huge Scale Japanese Association Thesaurus Construction. In Workshop Proceedings of the 22nd International Conference on Advanced Information Networking and Applications, AINA'08, GinoWan, Okinawa, Japan, March 25–28, 2008, pp. 1150–1155. IEEE Computer Society.
NAKAYAMA, K., HARA, T., AND S. NISHIO. [2007] A Thesaurus Construction Method from Large Scale Web Dictionaries. In Proceedings of the 21st IEEE International Conference on Advanced Information Networking and Applications, AINA'07, May 21–23, 2007, Niagara Falls, Canada, pp. 932–939. IEEE Computer Society.
NAKAYAMA, K., HARA, T., AND S. NISHIO. [2007] Wikipedia Mining for an Association Web Thesaurus Construction. In Proceedings of the 8th International Conference on Web Information Systems Engineering, WISE'07, Nancy, France, December 3–7, 2007, pp. 322–334. Lecture Notes in Computer Science 4831 Springer.
NASTASE, V. AND M. STRUBE. [2008] Decoding Wikipedia Categories for Knowledge Acquisition. To appear in Proceedings of the AAAI'08 Conference, Chicago, US.
NELKEN, R. AND E. YAMANGIL. [2008] Mining Wikipedia's Article Revision History for Traning Computattional Lingustic Algorithms. In Proceedings of the WIKI-AI: Wikipedia and AI Workshop at the AAAI'08 Conference, Chicago, US.
NGUYEN, D. P. T., MATSUO, Y., AND M. ISHIZUKA. [2007] Relation Extraction from Wikipedia Using Subtree Mining. In Proceedings of the AAAI'07 Conference, pp. 1414–1420, Vancouver, Canada, July 2007.
NGUYEN, D. P. T., MATSUO, Y., AND M. ISHIZUKA. [2007] Subtree Mining for Relation Extraction from Wikipedia. In Proceedings of the HLT-NAACL 2007, pp, 125–128.

NGUYEN, D. P. T., MATSUO, Y., AND M. ISHIZUKA. [2007] Exploiting Syntactic and Semantic Information for Relation Extraction from Wikipedia. In Proceedings of the IJCAI Workshop on Text-Mining and Link-Analysis, TextLink'07.

OLLIVIER, Y. UND P. SENELLART. [2007] Finding Related Pages Using Green Measures: An Illustration with Wikipedia. In Proceedings of the AAAI'07 Conference, pp. 1427–1433, Vancouver, Canada, July 2007.

OVERELL, S. E. AND S. RÜGER. [2007] Geographic co-occurrence as a tool for GIR. In Proceedings of the 4th ACM Workshop on Geographical Information Retrieval. Lisbon, Portugal.

OVERELL, S. E. AND S. RÜGER. [2006] Identifying and grounding descriptions of places. In Proceedings of the 3rd ACM workshop on Geographical Information Retrieval at SIGIR.

PEIRCE, C.S. [1877] The Fixation of Belief. Popular Science Monthly 12 (Nov. 1877), pp. 1–15.

PEI, M., NAKAYAMA, K., HARA, T. AND NISHIO, S. [2008] Constructing a Global Ontology by Concept Mapping using Wikipedia Thesaurus. In Proceedings of the 22nd International Conference on Advanced Information Networking and Applications, AINA'08, GinoWan, Okinawa, Japan, March 25–28, 2008, pp. 1205–1210. IEEE Computer Society.

PONZETTO, S. P. AND M. STRUBE. [2006]. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. In Proceedings of HLT-NAACL '06, pp.192–199.

PONZETTO, S. P. AND M. STRUBE. [2007a]. Knowledge Derived from Wikipedia for Computing Semantic Relatedness. Journal of Artificial Intelligence Research 30, pp. 181–212

PONZETTO, S. P. AND M. STRUBE. [2007b]. Deriving a Large Scale Taxonomy from Wikipedia. In *Proceedings of AAAI '07*, pp.1440–1445.

PONZETTO, S. P. AND M. STRUBE. [2007c]. An API for Measuring the Relatedness of Words in Wikipedia. In: Companion Volume of the Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 23–30 June, 2007, pp. 49–52.

PONZETTO, S. P. [2007] Creating a knowledge base from a collaboratively generated encyclopedia. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Doctoral Consortium, Rochester, NY, 22–27 April, 2007, pp. 9–12.

POTTHAST, M., STEIN, B., AND M. A. ANDERKA [2008] Wikipedia-Based Multilingual Retrieval Model. In Proceedings of the 30th European Conference on IR Research, ECIR'08, Glasgow.

POTTHAST, M. [2007] Wikipedia in the pocket: indexing technology for near-duplicate detection and high similarity search. In Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval.

QUINE, W.V.O. [1960] Word and Object. Cambridge, MA: MIT Press.

RANSDELL, J. [2003] The Relevance of Peircean Semiotic to Computational Intelligence augmentation. SEED Journal (Semiotics, Evolution, Energy, and Development).

RESNIK, P. [1999] Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 11, pp. 95–130.

RUBENSTEIN, H., AND J. GOODENOUGH. [1965] Contextual correlates of synonymy. Communications of the ACM 8(10), pp. 627–633.

RUIZ-CASADO, M., ALFONSECA, E., AND P. CASTELLS. [2005] Automatic assignment of Wikipedia Encyclopedic Entries to WordNet synsets. In Proceedings of AWIC'05.

RUIZ-CASADO, M., ALFONSECA, E., AND P. CASTELLS. [2007] Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. Data Knowledge and Engineering 61(3), pp. 484–499.

RUIZ-CASADO, M., ALFONSECA, E., AND P. CASTELLS. [2006] From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach. In Proceedings of the 1st International Workshop: SemWiki'06—From Wiki to Semantics. Co-located with the 3rd Annual European Semantic Web Conference ESWC'06 in Budva, Montenegro, June 12, 2006.

RUIZ-CASADO, M., ALFONSECA, E., AND P. CASTELLS. [2005] Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. In Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, NLDB'05, pp. 67–79, Alicante, Spain, June 15–17, 2005.

RUTHVEN, I. AND M. LALMAS. [2003] A survey on the use of relevance feedback for information access systems. Knowledge Engineering Review 18(2), pp. 95–145.

SCHOENHOFEN, P. [2006] Identifying Document Topics Using the Wikipedia Category Network. In Proceedings of the International Conference on Web Intelligence (IEEE/WIC/ACM WI'2006), Hong Kong.

SMITH, B., WILLIAMS, J., AND S. SCHULZE-KREMER. [2003]. The Ontology of the Gene Ontology. In Proceedings of AMIA Symposium, pp. 609–613.

SOON, W. M., NG, H. T., AND D. C. Y. LIM [2001]. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27(4), pp. 521–544.

SOWA, J. [2004]. The Challenge of Knowledge Soup. http://www.jfsowa.com/pubs/challenge.pdf.

STVILIA, B., TWIDALE, M. B., GASSER, L., AND L. SMITH. [2005]. Information Quality Discussions in Wikipedia. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Technical Report ISRN UIUCLIS–2005/2+CSCW

STRUBE, M. AND PONZETTO, S.P. [2006]. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In: AAAI '06, pp.1419–1424.

SUCHANEK, F. M., KASNECI, G., AND G. WEIKUM. [2007] Yago: a core of semantic knowledge. Proc 16th World Wide Web Conference, WWW'07. New York, NY: ACM Press.

SUCHANEK, F. M., KASNECI, G., AND G. WEIKUM. [forthcoming] Yago: A Large Ontology from Wikipedia and WordNet. Elsevier Journal of Web Semantics.

SUCHANEK, F. M, IFRIM, G., AND G. WEIKUM. [2006] Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents. In Proceedings of the Knowledge Discovery and Data Mining Conference, KDD'06.

SUH, S., HALPIN, H., AND E. KLEIN. [2006] Extracting Common Sense Knowledge from Wikipedia. In Proceedings of the ISWC'06 Workshop on Web Content Mining with Human Language technology.

SYED, Z., FININ, T., AND A. JOSHI. [2008] Wikipedia as an Ontology for Describing Documents. In Proceedings of the 2nd International Conference on Weblogs and Social Media, AAAI, March 31, 2008

THOM, A., PEHCEVSKI, J., AND A. M. VERCOUSTRE. [2007] Use of Wikipedia Categories in Entity Ranking. In Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia.

THOMAS, C.S., AND P. AMIT. [2006] Semantic Convergence of Wikipedia Articles. In Proceedings of the International Conference on Web Intelligence, IEEE/WIC/ACM WI'06, Hong Kong.

TORAL, A. AND R. MUÑOZH. [2007] Towards a Named Entity WordNet (NEWN). In Proceedings of the 6th International Conference on Recent Advances in Natural Language Processing, RANLP'07, Borovets, Bulgaria. pp. 604–608.

TORAL, A. AND R. MUÑOZH. [2006] A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In Proceedings of the Workshop on New Text at the 11th EACL'06. Trento, Italy.

TYERS, F. AND J. PIENAAR. [2008] Extracting bilingual word pairs from Wikipedia. In Proceedings of the SALTMIL Workshop at Language Resources and Evaluation Conference, LREC'08.

VERCOUSTRE, A. M., PEHCEVSKI, J., AND J. A. THOM [2007]. Using Wikipedia Categories and Links in Entity Ranking. In Pre-proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX'07, December 17, 2007.

VERCOUSTRE, A. M., THOM, J. A., AND J. PEHCEVSKI [2008] Entity Ranking in Wikipedia. In Proceedings of SAC'08, March 16–20, 2008, Fortaleza, Ceara, Brazil.

VIÉGAS, F.B., WATTENBERG, M., AND D. KUSHAL. [2004] Studying cooperation and conflict between authors with history flow visualizations. In Proceedings of SIGCHI'04, Vienna, Austria, pp. 575–582. New York, NY: ACM Press.

VIÉGAS, F., WATTENBERG, M., KRISS, J., AND F. VAN HAM. [2007] Talk before You Type: Coordination in Wikipedia. In Proceedings of the 40th Hawaii International Conference on System Sciences.

VÖLKEL, M., KRÖTZSCH, M., VRANDECIC, D., HALLER, H. AND R. STUDER. [2006] Semantic Wikipedia. In Proceedings of the 15th International Conference on World Wide Web, WWW'06, Edinburgh, Scotland, May 23–26, 2006.

VOORHEES, E. M. [1999] Natural Language Processing and Information Retrieval. In Pazienza, M. T. (editor) Information Extraction: Towards Scalable, Adaptable Systems, New York: Springer, pp. 32–48.

VOORHEES, E. M., AND HARMAN, D. [2000]. Overview of the eighth text retrieval conference (trec-8). In *TREC*, pp. 1–24.

VOSSEN, P., DIEZ-ORZAS, P., AND W. PETERS. [1997] The Multilingual Design of EuroWordNet. In Vossen, P. *et al.* (editors) Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12, 1997.

VRANDECIC, D., KRÖTZSCH, M., AND M. VÖLKEL. [2007] Wikipedia and the Semantic Web, Part II. In P. Ayers and N. Boalch (editors) Proceedings of the 2nd International Wikimedia Conference Wikimania'06. Wikimedia Foundation, Cambridge, MA, USA.

DE VRIES, A. P., THOM, J. A., VERCOUSTRE, A. M., CRASWELL, N., AND M. LALMAS. [2007] INEX 2007 Entity ranking track guidelines. In Workshop Pre-Proceedings of INEX 2007.

WANG, P., HU, J., ZENG H., CHEN, L., AND Z. CHEN. [2007] Improving Text Classification by Using Encyclopedia Knowledge. In Proceedings of the 7th IEEE International Conference on Data Mining, ICDM'07, 8–31 October 2007, pp.332–341.

WANG, G., ZHANG, H., WANG, H. AND Y. YU [2007a] Enhancing Relation Extraction by Eliciting Selectional Constraint Features from Wikipedia. In Proceedings of the Natural Language Processing and Information Systems Conference, pp. 329–340.

WANG, G., YU, Y., AND H. ZHU. [2007b] PORE: Positive-Only Relation Extraction from Wikipedia Text. In Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, ISWC/ASWC'07, Busan, South Korea.

WANG, Y., WANG, H., ZHU, H., AND Y. YU. [2007] Exploit Semantic Information for Category Annotation Recommendation in Wikipedia. In Proceedings of the Natural Language Processing and Information Systems Conference, pp. 48–60.

WATANABE, Y., ASAHARA, M., AND Y. A. MATSUMOTO. [2007] Graph-based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL.

WILKINSON, D.M., AND HUBERMAN, B.A. [2007] Cooperation and Quality in Wikipedia. In Proceeedings of the International Symposium on Wikis, pp. 157-164.

WU, F. AND D. WELD. [2007] Autonomously Semantifying Wikipedia. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM'07, Lisbon, Portugal, November 6–8, 2007, pp. 41–50.

WU, F. AND D. WELD. [2008] Automatically Refining the Wikipedia Infobox Ontology. In Proceedings of the 17th International World Wide Web Conference, WWW'08.

WU, F., HOFFMANN, R., AND D. WELD. [2008] Information Extraction from Wikipedia: Moving Down the Long Tail. In Proceedings of the 14th ACM SigKDD International Conference on Knowledge Discovery and Data Mining (KDD-08), Las Vegas, NV, August 24-27, 2008, pp. 635-644.

YANG, X. F. AND J. SU [2007] Coreference Resolution Using Semantic Relatedness Information from Automatically Discovered Patterns. In Proceedings of the 45th Annual meeting of the Association for Computational Linguistics, ACL'07, Prague, Czech Republic, pp. 528–535.

YANG, J., HAN, J., OH, I., AND M. KWAK. [2007] Using Wikipedia technology for topic maps design. In Proceedings of the ACM Southeast Regional Conference, pp. 106–110.

YU, J., THOM, J. A., AND A. TAM. [2007] Ontology evaluation using Wikipedia categories for browsing. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM'07, Lisbon, Portugal, November 6–8, 2007, pp. 223–232.

ZARAGOZA, H., RODE, H., MIKA, P., ATSERIAS, J., CIARAMITA, M., AND G. ATTARDI. [2007] Ranking Very Many Typed Entities on Wikipedia. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM'07, Lisbon, Portugal, November 6–8, 2007, pp. 1015–1018.

ZESCH, T. AND I. GUREVYCH. [2007] Analysis of the Wikipedia Category Graph for NLP Applications. In Proceedings of the TextGraphs-2 Workshop at the NAACL-HLT'07, pp. 1–8.

ZESCH, T., GUREVYCH, I., AND M. MÜHLHÄUSER. [2007] Comparing Wikipedia and German WordNet by Evaluating Semantic Relatedness on Multiple Datasets. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT'07, pp. 205–208.

ZESCH, T., GUREVYCH, I., AND M. MÜHLHÄUSER. [2008] Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In Proceedings of the Biannual Conference of the Society for Computational Linguistics and Language Technology, pp. 213–221.

ZLATIC, V., BOZICEVIC, M., STEFANCIC, H., AND M. DOMAZET. [2006] Wikipedias: Collaborative Web-based Encyclopedias as Complex Networks. Physical Review E, 74:016115.

ZIRN, C., NASTASE, V., AND M. STRUBE. [2008] Distinguishing between Instances and Classes in the Wikipedia Taxonomy. To appear in Proceedings of the ESWC'08.