Taking advantage of Wikipedia in Natural Language Processing

Tae Yano
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
taey@cs.cmu.edu

Moonyoung Kang
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
moonyoung@andrew.cmu.edu

Abstract

Wikipedia is an online encyclopedia created on the web by various participants. Although it is not created for the purpose of helping studies in language processing, its size and well-formed structure is attracting many researchers in the area. In this review, we selected five characteristic papers to show various creative uses of Wikipedia within the three years.

1 Introduction

Advancement of research in statistical natural language processing (NLP) critically relies on the corpora and their quality. The availability of high quality corpora is a major factor in deciding research trends in the field. Traditionally, corpus engineering has been viewed as a formidable undertaking involving massive resources and careful planning over several years; this kind of endeavor would be justifiable only if the reward of the labor is recognized by all involved parties. A common problem in such social projects is that, once completed, they tend to overprotect the invested paradigms, consequently slowing down the growth of new, unforeseen ones. Though new domains or tasks can be recognized at an individual level early on, the odiousness of corpus engineering could delay the growth of research for many years.

This standing norm of natural language research has been changing in recent years, as many high quality, voluminous, free on-line text collections become increasingly available. The critical difference in this new variety of "corpora" is that these electronic texts, unlike Penn Treebank and other traditional corpora, are not necessarily tailor-made for NLP research. Increasingly, the term "Corpus Engineering" takes on a different connotation; it is not about how they are made, but how they are put to use.

In recent years, Wikipedia has been gathering significant research interest in this context. Its ever-growing size, wide coverage, recency, and surprisingly high quality appear to be an appealing research opportunity. How to use this resource, however, is not fully understood. In this paper, we will review some of the recent studies that attempt to leverage this opportune resource in a creative manner. Though the types of NLP tasks vary, a common thread among the papers is their approach to using Wikipedia as a structured semantic knowledge base, capitalizing not only the meaning in isolation, but the relationships among them. We also opted for papers with research potential in a larger domain beyond Wikipedia. Therefore, we excluded papers dealing solely with the problems arising within Wikipedia (e.g., "how to encourage new users to participate", or "how to detect/correct erroneous entries"). In this paper, we focus on how traditional NLP tasks have advanced with the advent of Wikipedia.

The sections are divided as follows: Section 2 gives an overview of Wikipedia, highlighting its characteristics relevant to NLP research. section 3 gives a brief history of the interaction between

Wikipedia and NLP research through 2008, noting emerging trends. This section lays out the research landscape in the area and gives the historical context to the following literature reviews. We report five papers that we believe bear significance in this backdrop. We conclude with our view on these reviewed papers and future work in this area.

2 Wikipedia

2.1 Structure

Wikipedia is a multilingual, web-based, free content encyclopedia (Wikipedia:About) that attracts a lot of people. (Bunescu & Pasca, 2006) As of October 2008, it contained 11,389,385 articles in 264 languages, (Wikipedia:Size_of_Wikipedia) with more than 1.74 billion words and average of 435 words per article. (Wikipedia:Size_comparisons) English wikipedia is the largest and most well formed Wikipedia in terms of articles, words and the content of each article. English Wikipedia alone has more than 1 billion words, which is more than the 25 times the size of Britannica Encyclopedia. (Wikipedia:Statistics) What is more impressive about Wikipedia compared to off-line encyclopedias or corpora is its ability to grow or be updated by most recent news. According to Bunescu and Pasca (2006), often worthy news gets updated within a few days. For the past few years, English Wikipedia has grown at the speed of 500 thousand articles per year.

Wikipedia articles focus on explaining single natural human concepts, which provides a concrete text mirror of the abstract target concept. The concept articles are vertically integrated from broader to narrower concepts of the world. For example, the article about David Beckham has category links to 1975 births, living people, 1998 FIFA World Cup players, and 25 other broader concepts. Articles are also connected horizontally between same or similar concepts of different languages. The English article of David Beckham is linked to 55 pages of information about David Beckham in other languages.

As a natural language encyclopedia contributed to by many people, Wikipedia deals with problems of homonyms and synonyms. These problems are handled by disambiguation and redirection pages. A disambiguation page has list of links to unambiguous pages, usually in the order of popularity. For instance, page 'Beckham' has links to 'David Beckham', 'Brice Beckham', 'Victoria Beckham', and so on. A redirection page automatically redirects users to the dominant destination page. For example, 'Victoria Caroline Beckham' is redirected to 'Victoria Beckham'. Homonyms usually lead to a disambiguation page and synonyms usually lead to the same page by redirection.

Wikipedia article writers can avoid unnecessary redirection and disambiguation by making interwiki links directly to the unambiguous target page. A link in Wikipedia is formed by link and real content, in the form of [[link|real content]]. When 'link' is identical to 'real content', one can be omitted. So a typical Wikipedia sentence would look like this: '"Posh and Becks"' is the [[nickname]] for the [[England|English]] celebrity [[supercouple]] [[Victoria Beckham]] (formerly Victoria Adams and a member of the [[Spice Girls]] nicknamed "Posh Spice") and [[David Beckham]] (a leading [[soccer|footballer]]). This interwiki link structure can be exploited to figure out initials, abbreviations, and synonyms.

2.2 As a corpus

The internet provides a large, free, and easy to use corpus, but raw web data often contains ungrammatical text with lots of error. Wikipedia, whose aim is to be a general on-line encyclopedia tends to have better structured, well-formed, grammatical and meaningful, natural language sentences compared to ordinary internet text.

The official documents of governments, organizations, and newspapers provide a good source of well-written text. They are, however, off-line, often not freely available, not large enough, and in most cases, are domain limited. For example, Penn Treebank, or Wall Street Journal, is a good corpus of well-written English, but is biased towards financial English. Wikipedia, as an open encyclopedia, tries to represent all the natural concepts of human. Anybody can create a Wikipedia article about whatever they want and others can update each article. This makes

Wikipedia a source of knowledge following general distribution of human interest about the world. In addition each article is probabilistically less biased.

Following the convention of wiki, Wikipedia has acquired some suspicious critiques from researchers since the openness of Wikipedia may enable vandalism, bias, and errors of its content. (Denning et al. 2005; Riehle 2006; Kittur et al. 2007) However, Giles (2005) showed that the accuracy of Wikipedia actually rivals that of Britannica.

3 History

Wikipedia is a recent event in natural language research. Nonetheless, several distinctive trends have emerged in the course of its short history. The earliest of these trends is the use of Wikipedia as an external knowledge base for the question answering (QA) task such as in Ahn et al (2004). Most QA studies use Wikipedia in its intended way – as an encyclopedia to look up facts. In contrast, a large body of newer research using Wikipedia treats it as a naturally occurring text collection and focuses on how to "mine" some desired resource. One such resource is an annotated training corpus for supervised machine learning. Supervised learning, though it is an indispensable technique for NLP tasks, often suffers from a so called "resource bottleneck" (Mihalcea 2007). The performance of machine learning critically hinges on the training material, meanwhile relevant corpora are often ancient, meager, or out of domain. Wikipedia, though it is not meant to be an annotated corpus for any particular NLP task, appears as an impassable opportunity in this context. It is large, up-to date, and encompasses a wide variety of subjects. Other resource generation tasks conventionally viewed as unavoidable human labor, such as lexicon induction and feature engineering are another popular exploitation of Wikipedia. Multi-lingual NLP researchers are especially keen on utilizing Wikipedia since resource bottlenecks are often more of a problem than their monolingual counter parts.

The above mentioned desirable assets – size, coverage, recency – are not necessarily unique to Wikipedia among online text collections. Blogs, news groups, movie reviews, or QA sites and other on-line resources of collaborative nature, often possess those desirable assets as potential NLP resources. What makes Wikipedia singular among them is its aforementioned highly coherent structure, which makes automated analysis easier than free-text format resources. Though largely an unintended byproduct, Wikipedia seems to have hit a "sweet spot" for NLP researchers – it is flexible enough to be expansive, yet orderly enough to allow systematic analysis.

Taking advantage of its coherent structure to advance (any particular) NLP mission is the focus of a large body of Wikipedia related research. The tasks whose performance is thought to benefit from lexical semantic knowledge is most active in these fields: coreference resolution, word sense disambiguation, named entity categorization, and semantic role labeling. Interestingly, many such bodies of research share the common assumption that the structure of Wikipedia – its taxonomy, its hyperlinks – naturally encodes/reflects the semantic relationships in a language similar to human cognition since it was spontaneously developed by a mass. In a sense, structure of Wikipedia is not merely a convenience (that makes automated analysis possible) but is, itself, the heart of the matter. The challenge is in how to deal with such semantic crews in a principled manner.

4 Papers

In the following sections we review five papers that use Wikipedia as a structured semantic knowledge base in various degrees and in various forms (and awareness). The first two papers, Richman and Schone (2008) and Bunescu and Pasca (2006) are perhaps the most straight forward exploitations of the Wikipedia taxonomy. Nonetheless, the two papers are significant because they are some of the first to demonstrate the utility of Wikipedia in a non-trivial manner in tasks whose supervised paradigms tend to suffer from resource bottlenecks. They also provide early testimony of how "unintentional" corpora can be made to serve NLP tasks, and actually perform better in a realistic setting.

The following three papers take advantage of Wikipedia "structure" intentionally. Though they

are quite diverse, their approach to the problems can be summarized in the same way. They generalize Wikipedia's structure as a variation of a well-studied paradigm, and solve the problem utilizing the technique available in that pedagogy. Watanabe, Asahara, and Matsumoto (2007) generalized Wikipedia's categorical and hyper link structure into an undirected graph among the named entity (concepts) and semantic relationships between them. In addition, the structural learning technique was applied to classify those concepts into a set of known named entity classes. Gabrilovich and Markovitch (2007) applied a technique analogous to vector space analysis from information retrieval, introducing the algorithm in which a text is represented as a "bag of concepts" in lieu of "bag of topics" as latent semantic analysis. Strube and Ponzetto (2006) take the "Wikipedia as a graph" notion further, proposing its use as an instance of a "(lexical) semantic network" Steyvers and Tenenbaum (2005). Then they laid out its application to coreference resolution, using the wealth of theoretical underpinnings and techniques developed for more traditional semantic network, such as WordNet.

4.1 Richman & Schone: Mining Wiki Resources for Multilingual Named Entity Recognition

Named entity recognition (NER) refers to the isolation of named entity strings in text and the categorization of the strings into a predefined type. The two parts can be dealt with separately or jointly. Cucerzan (2007) pointed that in many area of NLP, including topic classification, machine translation, and information retrieval, NER has been recognized as an important task. There are many approaches to the task, but one popular approach is the use of gazetteers. "Gazetteer" is an umbrella term for some form of external resources, such as dictionaries, lookup tables, or compendia of rules. Gazetteers can be compounded manually or automatically. Another approach is to use machine learning, wherein the task is treated as a classification, or possibly ranking, of individual terms. The classification features are often gathered from (though not restricted to) the lexical and syntactical evidence surrounding the query term (the term under consideration). Note that both approaches may require a non-trivial degree of human involvement in the engineering cycle for realistic application, where realistic issues such as emergence of new entities, or adaptation to a new domain frequently require generation and maintenance of gazetteers or the annotation of new training corpus for machine learning.

The macro level goal of the paper by Richman and Schone (2008) is to alleviate this resource bottleneck in NER engineering, (i.e. how can entity recognition and categorization be approached with as little human involvement as possible?). Their specific goal here is to demonstrate how to exploit Wikipedia to this end.

Richman and Schone (2008) shows one of the simplest results possible with the structure of Wikipedia. The goal of the paper is to automatically generate huge, multilingual tagged data for NER using information only from Wikipedia with no linguistic, foreign language, or outer knowledge, like WordNet.

As mentioned, Wikipedia articles have a lot of internal links to other Wikipedia articles. Since each article is a concept, each interwiki link can be considered a possible named entity, and a named entity tagged text can be easily generated.

To tag each named entity with proper named entity type, the authors exploited Wikipedia structure. For a non-English concept, the English page of the identical concept was searched. If there was no identical English concept, the English page of broader concept was searched. Once a matching English page was found, the category of that page was used to determine the type of named entity using a simple rule-based method of English. For example, a category whose name contains a string 'people by' was considered as a person named entity type. When multiple categories match different tag types, the desired tag is selected through a majority vote. When no matching categories are found, the page is checked to see if it is a disambiguation page and is tagged by the dominant type. The tagging result is rechecked by Wiktionary, an online collaborative dictionary and revised, if necessary.

Beyond the interwiki link, authors include more tricks to increase tagged data in their corpus. The authors tagged text of two to four words, which partially match Wikipedia concept. They also

include language independent entities like money, date, time, percent, and other special case tricks. For example, the "X.X." prefix was used to determine the following string a named entity of person type.

The authors generated tagged data in Spanish, French, Ukrainian and three other languages. They used PhoenixIDF, their modified version of BBN's IdentiFinder from Bikel et al. (1999), to train and test the tagged data. For test data, newswire articles were used.

	Newswire Accuracy	Size of annotated corpus for similar result
Spanish	.827	20,000~40,000
French	.847	40,000
Ukranian	.747	15,000~20,000

Table 1: test result

Spanish corpus reached f-measure of .827, which is about the result attainable from a corpus of 20,000~40,000 words of human annotated data. [table 1] Other languages show various results rivaling a human annotated corpus of 15,000~40,000 words.

The paper is unique in the sense that it uses wiki structure to generate multilingual corpus without any linguistic or foreign language knowledge. Without human annotation, it automatically produces tagged corpus comparable to the large human annotated corpus. However, most procedures were rule-based and some linguistic, language dependent tricks were used to improve the result. Even so, the paper illustrates a good way of exploiting Wikipedia structure.

4.2 Bunescu & Pasca: Using Encyclopedic Knowledge for Named Entity Disambiguation

Like Richman and Schone, the goal of the paper by Bunescu and Pasca is to demonstrate how to exploit Wikipedia to alleviate resource bottleneck in a similar task, named entity disambiguation. In doing so, the authors' intention is not only taking advantage of Wikipedia's syntax for automatic analysis, but also of the purported semantic relationships encoded in the collaborative taxonomy structure of the Wikipedia articles. The underlying assumption here is that the structure of Wikipedia lends itself to a set of useful clues for identifying entity classes, a system that leverages such assets to perform better.

Given a term to disambiguate its sense in some context and a massive on-line encyclopedia with many entity entries, one possible solution is to measure the similarity between the term's surrounding context and each entry to which the term can possibly refer. In other words, named entity recognition can be cast as a ranking problem among the candidate entries which is the basic idea with which Bunescu and Pasca started. Using Wikipedia, candidate entries can be gathered from 'disambiguation pages', where all possible senses of a term are listed on one page. Similarly, score can be computed with the standard tf-idf such as the following:

$$\hat{e} = arg\max_{e_k} score(q, e_k)$$

where q is a query term to be disambiguated. e_k is a candidate wiki entry, and $\hat{e}$ is the output from the system, the prediction on which wiki entry (which sense) the query term refers to. Taking the standard tf-idf cosine similarity scheme, the scoring function is the following:

$$score(q, e_k) = cos(q.T, e_k.T) = \frac{q.T}{\|q.T\|} \frac{e_k.T}{\|e_k.T\|}$$

where q.T is e_k.T is the standard vector representation of the words in the query context and the words in the candidate article, respectively.

The authors empirically found that this simple method often performs sub-optimally for two reasons: 1) The encyclopedia entry may be too short or incomplete. 2) The word usage between the articles and the text surrounding the term may be different. To overcome these problems, the authors propose a similarity scoring function that considers not only the similarity between the context words (the words which surround the query term) and the candidate article itself, but also the "correlations" between those words and all the Wikipedia categories to which the article belongs. As noted, all entries appearing in Wikipedia belong to at least one category. Categories, in turn, belong to more general categories. The authors' intention here is to take advantage of the nested categorical structure for making a more informed similarity measurement. Consider a query term "John Williams", a composer. It is likely to appear in the

context including words such as "conducts" or "Star Wars". Those words may or may not appear in the wiki entry for "John William (composer)", but its categories, "Film composer", "Composer", or "Musician" most likely has a high correlation with the words. Based on this observation, the authors proposed to augment the tf-idf cosine similarity scoring in the following way:

$$\hat{e} = arg\max_{e_k} \mathbf{w}\ \Phi(q, e_k)$$

$$\phi_{cos}(q, e_k) = cos(q.T, e_k.T) \quad (4)$$

$$\phi_{w,c}(q, e_k) = \begin{cases} 1 & \text{if } w \in q.T \text{ and } c \in e_k.C, \\ 0 & \text{otherwise.} \end{cases}$$

The feature vectors $\Phi(q,e_k)$, consists of two kinds of feature functions: A cosine similarity scoring function, and V*C (V= size of vocabulary and C= the total number of wiki categories) many delta functions, each of which indicates co-occurrence of a wiki category and a context word. The weight vector, W, represents the magnitude of correlation between each word and category.

Having formed the classification task as shown, the remaining task is to lean the weight vectors. The specific learning algorithm the authors chose for the experiment is support vector machine (SVM) light by Joachims (1999). As for the training corpus, interestingly, the authors exploit another feature of Wikipedia. To compose the set of annotated data, they harvest the local inter-wiki hyper links (which unambiguously link to an article) in articles and the surrounding text, circumventing the human annotation. They gathered 1,783,868 annotated examples in this manner.

In their evaluation, the authors compared simple cosine similarity measurement and the proposed scheme that considers category-word correlation. The proposed model was trained with the portion of the annotated examples. Furthermore, they experimented with various sets of features for the proposed model because including all the Wikipedia categories is too computationally intensive. In this paper, they conducted four of such variations. In each experiment, the classification accuracy on test examples of two schemes are compared. The authors reported that the proposed method significantly outperforms the cosine similarity method in all four experiments.

In essence, Bunescu and Pasca demonstrated two important points in this paper: One is the usability of Wikipedia as a source of gazetteer and the second is the use of annotated data. Their dual strategy is in a sense the combination of the gazetteer and machine learning approach, and in both parts Wikipedia is used cleverly to circumvent potentially costly human labor. The other point demonstrated is the exploitation of Wikipedia categories for the disambiguation task. As their evaluation showed, the taxonomical features coaxed from Wikipedia's categorical structure improve performance, even though the categorization of those articles was largely done in a spontaneous manner based solely on human intuition without compulsory guidelines. Although this paper lacks the performance evaluation against the existing NER strategy, thus their competence in a wider world is unclear, clever use of Wikipedia brings considerable saving in human labor. One problem not addressed is the issue of scalability. In this paper, the authors were compelled to shrink both the training data set and the number of features considerably to conduct their experiments in a manageable time scale. By default, the number of features is the size of vocabulary times the number of Wikipedia category. Learning the weight vector for all those features is computationally very expensive. Be it the algorithm selection or feature selection, the care to cope with the curse of dimensionality seems to be necessary, of which the authors left for future explorations. As we see in other papers we report in this review, the problem of scalability is a recurring theme in research dealing with Wikipedia.

4.3 Watanabe, et al: A Graph-based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields

Conditional random field (CRF) (Lafferty, et al, 2001) is an undirected, graphical model particularly well suited to do structural prediction In the prediction, the iid assumption among the random variables does not necessarily hold. Tang et al. (2006) and Zhu et al. (2006) used this technique to model structured document collections, such as a collection of hyperlinked texts, and several hyperlink specific extensions to the CRF has been proposed. In this paper,

Watanabe and his colleagues extend this idea to Wikipedia taxonomy to solve named entity categorization problem.

The originality of this literature is, in the way the authors generalized, named entity classification as a text classification of Wikipedia articles. Each article in Wikipedia is taken as a distinctive "concept", which has (or will have) its unique associated (named entity) label. The task is to discover true labels for articles. Furthermore, they generalized the problem as a structural labeling task. Given a group of articles to categorize, one possible approach is to classify each one without considering the labeling for others. This approach would be reasonable if there are no relationships between the articles. The authors presume that this is not the case among Wikipedia articles. some articles are more closely tied together than others, thus the task would be better approached as a sequence labeling problem. A labeling decision can then influence other, related articles' labeling.

Specifically, the authors proposed the following classification scheme: First, Wikipedia articles and relationships between them are (somehow) abstracted into a set of cliques. Then, a conditional probability model (graph-base CRF) over the class label sequence (= named entity category) in a form of exponential model is drawn. Each clique forms a potential function, and potential functions are factorized according to the set of local features. Features here include the title of articles, heading words, articles' categories, and labels for other articles within the clique. The weights on each potential function over the cliques are estimated using quasi-newton methods with gaussian prior. Since the proposed model may include loops, Tree-based Reparameterization in Wainwright et al (2003) is used for approximating inference. Categorization of new sets of articles are made by computing the marginal probability via maximum a posteriori (MAP) inference.

Though there are several important aspects to the success to this approach, the authors focused on how to abstract the relationships among Wikipedia articles. There are many possible "relationships" among articles. Ideally, one would want to capture relationships that help the task at hand and leave out others. Meanwhile, in dealing with structural prediction, the performance (in both optimality and efficiency) often depends on the complexity of the underlying structures. Linear chains or trees are much easier than a general graph, although the latter has more expressive power. Trade-off between these concerns is important since it directly relate to the issue of scalability.

In their paper, the authors experimented with encoding three types of relationships into cliques: sibling, cousin, and relatives. Although it is conceivable to extend the idea into more general Wikipedia categorization structure or hyper-linkage structure, in this report the authors used only the "List of ..." pages to find those relationships. Given a set of Wikipedia pages that included the list constructions, the article cliques were constructed in the following manner: First, intrawiki anchor texts (which links to an article inside Wikipedia) are extracted from the page. These form the set of nodes in the graph. Edges are drawn between these nodes when they meet the following definition of the relationships: 1) A "sibling" is a edge between the nodes who have a common parent in the HTML table structure. 2) A "cousin" is a edge between the nodes who has a common grand parent node, and occupy the same position in their sibling ordering. 3) A "relative" is a edge between a node and another who is a sibling of the first node's ancestor.

The intuition here is that those articles appearing in the same or nearby positions in the list hierarchy are often in the same or related named entity class. By encoding the topological relationships into the clique, a particular assignment of a label to one class will influence the assignment to the other.

In their evaluation, the authors aim to investigate two points: if the relationship information helps the classification task, and which cliques (or combination of them) would help the task the most. To this end, authors experimented with several variations of CRF with different combinations of edge types (essentially different types of potential functions), additionally the maximum entropy (MaxEnt) and support vector model (SVM) classifier trained with the same data set minus the edge features was explored. The training data set consisting of 2,300 articles was extracted from Wikipedia in the aforementioned manner. Sixteen thousand anchor texts were

identified as a named entity, and were hand-annotated. The number of cliques varied depending on the relationships encoded, ranging from 876 (cliques which include all types of edges) to 16,000 (no clique considered).

The authors report that categorization using CRF outperforms the non-structural classifier in many cases, though it scores lower for the rarer types of named entities (such as EVENT, or UNIT). They also found that the training time is considerably longer for CRF. Using full set of edge types, the training time is as much as ten times more than with plain SVM. This is largely due to increased complexity of the cliques. As much as 36% of the training examples with full edge set contained at least one cycle. The comparison between various combinations of cliques demonstrates that all edges are not equally important. Some help performance, and others introduce more loops into the graph (making approximation more costly). Naturally, the more complicated the graph, the higher its probability of having loopy construction. The experimental result shows that the best performance is achieved by those configurations that include cousin relationships and relative relationships. The results also show that cousin relationships tend to bring larger single performance gain.

4.4 Strube & Ponzetto: WikiRelate! Computing Semantic Relatedness Using Wikipedia.

A significance of Watanabe paper in terms of Wikipedia mining is that it manages to realize an abstract "Wikipedia as a graph" notion within a mathematically principled structural learning paradigm. The following paper by Strube and Ponzetto also takes the same inspiration, but formalized in yet another pedagogy, lexical semantic networks.

Lexical semantic networks have been an active research paradigm in computational linguistics for many years. In earlier studies, researchers' interests are more focused on how such elusive notions as "semantic" can be formalized in a rigorous manner. For NLP researchers, the issue is as much a practical concern as a philosophical interest since the codification of "semantic relations" would undoubtably benefit many NLP tasks.

In its simplest (and the most practical) definition, semantic relatedness indicates how much two concepts are related in a taxonomy by using "relations" between them (i.e. hyponymic or hyperemic, meronymic, and other functional relations such as is-a, or has-part). When limited to hyponymy or hyperonymy, the measure quantifies semantic similarity.

Many approaches in quantifying such inherently qualitative aspects of language are proposed by NLP researchers to date. One interesting approach is to codify lexical resources into a network of semantic concepts, or a lexical semantic network. This is an attractive method because by generalizing the semantic relations into a network, one can utilize a wealth of principled formalism developed for graph analysis. The trend has gained momentum in recent years due to the advancement in large-scale network analysis motivated by search engine technology. The results of this technology is a succession of research applicable to any collection of data generalizable to a network of random variables, including lexical semantics. Another, perhaps more important advancement, is the advent of WordNet (Fellbaum, 1998), which made it possible to put semantic theories in realistic test and develop usable applications. For this reason, WordNet is the de facto data set for semantic relations research.

Although WordNet is an excellent resource, its limited coverage becomes problematic in engineering realistic applications. For example, WordNet 2.1 does not include information about named entities such as "Condoleezza Rice", or "The Rolling Stones", as well as the specialized concept such as "Semantic Network". Meanwhile, Wikipedia, though it is less formal, has excellent size and coverage. This is the motivation behind Strube and Ponzetto's study on semantic relatedness using Wikipedia.

Previous to Strube and Ponzetto, Zesch and Gurevych (2006) conducted a series of graph theoretic analyses on Wikipedia's taxonomy structure and found it to be "as scale-free, small world graph like other well-known lexical semantic networks". An important implication in

this is that Wikipedia, on principle at least, can be used for the NLP tasks in the similar manner as other traditional semantic networks. Those findings withstanding, there are easily foreseeable problems in using Wikipedia as a lexical semantic network. For one, there is a question of consistency; unlike other lexical semantic resources, Wikipedia is created by voluntary collaborations and thus, is not perfectly organized. There is also a concern of its size, which is a magnitude bigger than any other artificial semantic network.

The macro level goal of Strube and Ponzetto's paper is to attest to the viability of Wikipedia as a lexical resource in this context. The authors conducted several semantic relatedness experiments using methods developed mostly for WordNet (Budanitsky and Hirst, 2006). Discussed here are three types of similarity measurements: 1) path based measure, 2) information content base measure, and 3) text overlap based measurement. The first two are based on graph analysis. The basic path based measurement computes the similarity as a function of edge count (distance in the graph), though there are many more sophisticated varieties. The original information content method in Resnik (1995) computes the score based on the probability of two words occurring together in a corpus, measuring "extent to which they [the concepts] share information. Seco et al. (2004) proposed a variety incorporating the informativeness of words' parent categories. The third type of similarity measurement counts the overlapping words in the context the words appear (in the case of current research, the count of overlapping words appeared in the Wikipedia articles.)

The authors chose two varieties of each of the three types of methods and conducted the evaluation on several standard test data sets for similarity scoring using Wikipedia and Wordnet. They found Wikipedia to be a competitive alternative to WordNet when the data set is large, suggesting Wikipedia may have advantage in realistic settings. The authors also conducted coreference resolution tasks using those similarity measurement, adding them as features for the standard, supervised classifier. The results show that the classifier using both Wikipedia and WordNet similarity measurement performs the best.

The experiments reported here are seemingly modest. Semantic similarity measurement is one of the most well studied methods in lexical semantic analysis and one of the easiest to apply to Wikipedia. Meanwhile, Mihalcea (2005) and Navigli and Lapata (2007) used synthetic semantic networks such as WordNet for far more sophisticated graph-based algorithms in large scale tasks. To be sure, giving the greater constitutional dissimilarity between Wikipedia and Wordnet (such as its folksonomical organization and efficiency issues due to its size), the work reported here is valuable. Whether Wikipedia is a semantic resource of the same caliber as WordNet, however, is not yet fully tested.

In the following paper, Gabrilovich and Markovitch take completely different approach in Wikipedia exploitation for the same task, word similarity measurement. In doing so, they focus on the unique assets of Wikipedia not found in WorldNet, the folksonomical, organic origin of Wikipedia ontolology. To the authors, those aspects are not drawbacks to be overcome, but competitive edges that make Wikipedia better for harvesting “natural human concepts”.

4.5 Gabrilovich & Markovitch: Computing semantic relatedness using Wikipedia-based explicit semantic analysis

Explicit semantic analysis (ESA) is a novel and unique way of representing the meaning of arbitrary length text, representing text as a weighted vector of natural human concepts. ESA-Wikipedia uses Wikipedia concepts as its base orthonormal concepts.

Latent semantic analysis (LSA) by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) describes the meaning of the text by a term-document matrix whose row corresponds to term and column to document. Each element of the matrix represents the relatedness of term and document, usually by tf-idf score. The matrix is then decomposed and shrunk into relations of terms and concept, concept and concept, and concept and document. While converting the matrix, semantically unrelated terms are often combined together, making LSA concept linguistically uninterpretable or latent.

ESA-Wikipedia has overcome this problem of latent concept by using the explicit concept of Wikipedia. Each article of Wikipedia is a page directly describing the concept of the title. Each Wikipedia concept is represented by a weighted sum of terms, using tf-idf score like in LSA. From this weight, inverted index describing each term by concepts can be calculated.

The way of computing and comparing weighted vectors is similar to LSA. For any text T, T can be represented by weighted sum of its word tokens vectors, $T = \Sigma_i(w_i * v_i)$, where w_i is the word token and vi is the corresponding weight, (i=1~N). Let cj each concept in Wikipedia (j=1~M), and k_{ij} inverted index of word token w_i by c_j. Then, $w_i = \Sigma_j(k_{ij} * c_j)$. Therefore, the whole text T becomes $\Sigma_i(v_i * \Sigma_j(k_{ij} * c_j)) = \Sigma_j(\Sigma_i(v_i * k_{ij}) * c_j)$.

Similarity of a text and a concept is calculated by cosine value, $\cos\theta = (T \circ c_j)/(|T||c_j|)$. Since |T| is constant over all concepts and all the concepts are orthonormal, order of the relatedness of each concept to the text can simply be decided by comparing their weights, $\Sigma_i(v_i * k_{ij})$. By ordering concepts by weight, concepts closest to the term can be found easily. For example, by looking at the list of ten closest concepts of 'Bank of America' and 'Bank of Amazon' we can easily see that what ESA thinks about each text. [Table 2]

#	Ambiguous word: "Bank"	
	"Bank of America"	"Bank of Amazon"
1	Bank	Amazon River
2	Bank of America	Amazon Basin
3	Bank of America Plaza (Atlanta)	Amazon Rainforest
4	Bank of America Plaza (Dallas)	Amazon.com
5	MBNA	Rainforest
6	VISA (credit card)	Atlantic Ocean
7	Bank of America Tower,	Brazil
8	New York City	Loreto Region
9	NASDAQ	River
10	MasterCard	Economy of Brazil
	Bank of America Corporate Center	

Table 2: First ten concepts of the interpretation vectors for texts with ambiguous words

For evaluation of ESA-Wikipedia, Gabrilovich & Markovitch built ESA-ODP(open directory project, http://www.dmoz.org). ODP is a large category site of webpages. ESA-ODP used categories of ODP as the concept and text of webpages of each category's link as the text of the concept. For baseline system WordNet, LSA, and WikiRelate! by Strube and Ponzetto (2006) were used. WordSimilar-353, human annotated corpus of relatedness of 353 words by FinkelStein et al., (2002), were used to test word relatedness prediction. Australian Broadcasting Corporation's news mail service by Lee et al., (2005) was used to test text relatedness prediction. ESA techniques achieved substantial improvements over prior studies and ESA-Wikipedia showed the best result. [Table 3, 4]

Algorithm	Correlation with Humans
WordNet (Jarmasz, 2003)	0.33-0.35
Roget's Thesaurus (Jarmasz, 2003]	0.55
LSA (Finkelstein et al., 2002)	0.56
WikiRelate! (Strube and Ponzetto, 2006)	0.19-0.48
ESA-Wikipedia	0.75
ESA-ODP	0.65

Table 3: Computing word relatedness

Algorithm	Correlation with Humans
Bag of words (Lee *et al.*, 2005)	0.1-0.5
LSA (Findelstein et al., 2002]	0.60
ESA-Wikipedia	0.72
ESA-ODP	0.69

Table 4: Computing text relatedness

ESA is unique in the sense that its semantic representation is based on explicit natural human concept about the world. Wikipedia, whose articles are well formed toward a concept, seems to be the best corpus to train ESA. Unlike LSA, ESA-Wikipedia uses knowledge of the world over statistical methods. Unlike WordNet, ESA-Wikipedia can be created without any expert. Its data is still growing and is much better at handling ambiguous words statistically using related terms. Above all, ESA-Wikipedia is easy to understand by humans since it is based on human concepts.

5 Conclusion

Wikipedia is a promising resouce for NLP. It is a large, well-formed, human concept oriented, and open domain. Although Wikipedia is not tailored for NLP, research based on Wikipedia is showing promising results.

Richman and Schone (2008), and Bunescu and Pasca (2006) showed that automatically generated tagged corpus' from Wikipedia can substitute the human annotated corpus for tasks. Watanabe et al. (2007), Strube and Ponzetto (2006), Markovich and Gabrilovitch (2007) showed that the structure of Wikipedia can be exploited in ways beyond creating tagged corpus.

The results of these studies will be more relevant in future since the legacy of Wikipedia is keep growing. Articles in Wikipedia are constantly being added and updated. Knowledge resources created by wiki software are becoming more and more common, which operate on the same folksonomic principle that made Wikipedia be one of the most variable resources on the web. Moreover, there are some concerted efforts in making Wikipedia a better semantic resource in general. Volkel et al (2007) suggested adding more formal codification of interrelation between the concepts (articles) to the grammar of Wikipedia, which enables precise semantic understanding of content by machine. Although the studies introduced here have not beaten the state-of-art performance on any given task, the significance of them is in pioneering attempts. Considering the great promises in the development of Wikipedia, we strongly believe that these studies' contributions to NLP should be recognized.

References

Ahn, D., Jijkoun, V., Mishne, G., Muller, K., de Rijke, M., and Schlobachm S. 2004. *Using Wikipedia at the TREC QA track.*In proc. The Thirteenth Text Retrieval Conference (TREC 2004)

Bikel, D., R. Schwartz, and R. Weischedel. 1999. *An algorithm that learns what's in a name.* Machine Learning, 211-31

Budanisky, A & G. Hirst. 2006. *Evaluating WordNet-based measurement of semantic distance*. Computational Linguistics, 32(1)

Bunescu, Razvan. and Pasca, Marius., 2006. *Using Encyclopedic Knowledge for Named Entity Disambiguation* In proceedings of 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-06)

Cucerzan, Silviu., 2007. *Large-Scale Named Entity Disambiguation Based on Wikipedia Data.* In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-07)

Denning, P.; Horning, J.; Parnas, D.; and Weinstein, L. 2005. *Wikipedia risks.* Commun. ACM 48(12):152–152.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. 1990. *Indexing by latent semantic analysis.* JASIS, 41(6):391–407, 1990.

Fellbaum, C. (Ed.). 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

Finkelstein, Lev., Gabrilovich, Evgeniy., Matias, Yossi., Rivlin, Ehud., Solan, Zach., Wolfman, Gadi., and Ruppin, Eytan. *Placing search in context: The concept revisited.* ACM TOIS, 20(1):116–131, January 2002.

Gabrilovich, Evgeniy., Markovitch, Shaul., 2007. *Computing semantic relatedness using Wikipedia-based explicit semantic analysis.* In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI07)

Giles, J. 2005. *Internet encyclopaedias go head to head.* Nature 438:900–901.

Jarmasz, Mario. 2003. *Roget's thesaurus as a lexical resource for natural language processing.* Master's thesis, University of Ottawa.

Joachims, Thorsten. 1999. *Making Large-Scale SVM Learning Practical.* In Advances in Dernel Methods - Support Vector Learning Pages 169 - 184. MIT Press.

Kittur, A.; Suh, B.; Pendleton, B. A.; and Chi, E. H. 2007. *He says, she says: conflict and coordination in wikipedia.* In CHI, 453–462.

Lafferty, John., McCallum, Andrew., and Pereira, Fernando. 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data* In Proc. The Eighteenth International Conference on Machine Learning. (ICML-01)

Lee, Michael D., Pincombe, Brandon., and Welsh, Matthew. *An empirical evaluation of models of text document similarity.* In CogSci2005, pages 1254–1259, 2005.

Mihalcea, Rada. 2005. *Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling.* In Proc. Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP-05).

Mihalcea, Rada. 2007. *Using wikipedia for Automatic Word Sense Disambiguation.* In Proc. The Annual Conference of the North American Chapter of the

Association for Computational Linguistics (NAACL-HLT 2007)

Navigli, Roverto and Lapata, Mirella. 2007. *Graph Connectivity Measures for Unsupervised Word Sense Disambiguation.* In Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI-07)

Resnik, P. 1995. *Using Information Content to Evaluate Semantic similarity in a Taxonomy*. In Proc. the Fourteenth International Joint Conference on Artificial Intelligence (IJCAL-95)

Riehle, D. 2006. *How and why wikipedia works: an interview with angela beesley, elisabeth bauer, and kizu naoko.* In Proceedings of the International Symposium on Wikis, 3–8.

Richman., Alexander E.; Schone., Patrick., 2008., *Mining Wiki Resources for Multilingual Named Entity Recognition* In proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08)

Seco, N., T. Veale & J. Hayes. 2004. *An Intrinsic Information Content Metric for Semantic Similarity in WordNet.* In Proc. 16th European conference on artificial intelligence (ECAI-04)

Strube, Michael. and Ponzetto, Simon Paolo. 2006. *WikiRelate! Computing Semantic Relatedness Using Wikipedia.* The Twenty-First National Conference on Artificial Intelligence (AAAI-06)

Steyvers, Mark., and Tenenbaum, Josh., 2001. *Small Worlds in Semantic Networks.* Unpublished Manuscript, Department of Psycology, Stanford University, 2001

Tang, Juanzi Li Jie., Hong, Mingcai., and Liang, Bangyong. 2006. *Tree-Structured Conditional Random Fields for Semantic Annotation.* In proc The 5th International Semantic Web Conference (ISWC2006)

Völkel, Max., Krötzsch, Markus., Vrandecic, Denny., Haller, Heiko., and Studer, Rudi. 2007. *Semantic Wikipedia.* Proceedings of the 15th international conference on World Wide Web

Wainwright, Martin., Jaakkola, Tommi., and Willsky, Alan. 2003. *Tree-Based Parametrization Framework for Analysis of Sum-Product and Related Algorithm.* IEEE Transactions and Information Theory, 45(9) Page 1120 - 1146

Watanabe, Yotaro., Asahara, Masayuki., and Matsumoto, Yoji., 2007. *A Graph-based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields.* In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-07)

Zesch, Torsten. and Gurevych, Iryna. 2007. *Analysis of the Wikipedia Category Graph for NLP Applications* In Proc. The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007), Textgraph-2

Zhu, Jun., Nie, Zaiqing., Wen, Ji-Rong ., Zhang, Bo., and Ma, Wei-Ying. 2006. *Simultaneous Record Detection and Attribute Labeling in Web Data Extraction.* In proc The Twelfth *ACM SIGKDD* International Conference on Knowledge Discovery and Data Mining (ACM-SIGKDD-06)

Wikipedia:About. *http://en.wikipedia.org/wiki/Wikipedia:About*

Wikipedia:Size_comparisons. http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

Wikipedia:Size_of_Wikipedia. http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Wikipedia:Statistics. http://en.wikipedia.org/wiki/Wikipedia:Statistics#Automatically_updated_statistics