WikiSense: Supersense Tagging of Wikipedia Named Entities Based WordNet

Joseph Chang[a], Richard Tzong-Han Tsai[a], and Jason S. Chang[b]

[a]Department of Computer Science and Engineering, Yuan Ze University
No. 135 Yuan-Tung Road, Chungli, Taoyuan, Taiwan 32003
{s951533, thtsai}@mail.yzu.edu.tw

[b]Department of Computer Science, National Tsing Hua University
No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013
jschang@cs.nthu.edu.tw

Abstract. In this paper, we introduce a minimally supervised method for learning to classify named-entity titles in a given encyclopedia into broad semantic categories in an existing ontology. Our main idea involves using overlapping entries in the encyclopedia and ontology and a small set of 30 handed tagged parenthetic explanations to automatically generate the training data. The proposed method involves automatically recognizing whether a title is a named entity, automatically generating two sets of training data, and automatically building a classification model for training a classification model based on textual and non-textual features. We present *WikiSense*, an implementation of the proposed method for extending the named entity coverage of WordNet by sense tagging Wikipedia titles. Experimental results show *WikiSense* achieves accuracy of over 95% and near 80% applicability for all NE titles in Wikipedia. *WikiSense* cleanly produces over 1.2 million of NEs tagged with broad categories, based on the lexicographers' files of WordNet, effectively extending WordNet to form a very large scale semantic category, a potentially useful resource for many natural language related tasks.

Keywords: semantic category, word sense disambiguation, WordNet, Wikipedia

1 Introduction

Machine readable natural language resources, like ontologies or semantic categories, are crucial in natural language processing or information retrieval tasks that demands world knowledge. Such tasks include question classification and answering, knowledge mining and semantic search. In these tasks, lexical databases such as WordNet or Suggested Upper Merged Ontology (SUMO) are widely used to provide meaning representation and semantic relations. These handcrafted ontologies rely on manual compilation and maintenance of small groups of experts over long periods of time. New words and phrases are added to the vocabulary progressively over the years through each new release. With the scale close to that of a dictionary, these resources consists mostly of common nouns, verbs, adjectives, adverbs, and a small amount of named entities. Only highly well-known people, places, organizations, and events, are included. Although these resources are very useful for many natural language tasks, they sometimes suffer severely from the out-of-vocabulary problem due to limited coverage, most notably for lack of NEs. Furthermore, for NEs they cover, there are also issues with consistency and relevance. For example, *Charles Dickens* is found in WordNet, but none of the books he wrote (e.g., *A Tale of Two Cities*). The generally accepted lexicographer guidelines called for using the powerful tool of frequency to make informed linguistic decisions about creating and framing an entry. A typical Web search (e.g., Google, http://www.google.com) show *Celine Dion,* a contemporary Canadian pop singer, has slightly higher visibility on the Web than *Johnny Cash*, the late American country singer. (13,100,000 v.s. 11,600,000 page counts) However, *Celine Dion* is not listed in WordNet, while *Johnny Cash* is listed as an instance of singer#n#1. Such

inconsistent and poor coverage could be problematic for building robust language systems for practical applications.

In effort to automatically produce broad and general semantic categories of NEs, we turn to Wikipedia, an online encyclopedia contributed by millions of volunteers all around the world. Wikipedia consists millions of articles of all kinds, including rich information of constantly emerging NEs not found in existing dictionaries and ontologies. Owing to the tremendously active participation from its contributor community, Wikipedia is constantly infused with the newly created words and phrases, especially NEs, such as names of movies, books, celebrities, and events, both new and old. For instance, Wikipedia includes all four NEs mentioned above: *Charles Dickens*, *A Tale of Two Cities*, *Johnny Cash*, and *Celine Dion*. Intuitively, to extend WordNet, a feasible approach is to classify Wikipedia NE titles into semantic categories in WordNet, thus greatly extending WordNet's coverage of NEs.

Table 1: Simplified concept of WikiSense classification process.

Input	Wikipedia entry: Celine Dion	
Features	genus	genus:actress, genus:singer-songwriter, genus-lexfile:noun.person, genus-lexifile:noun.person,
	category	cat:celine_dion, cat:canadian_singers, cat:canadian_female_singers, ... wncat:singer#n#1, wncat:entertainer#n#1, ...
	pronoun	dominating-pronoun:PERSON
Output	Celine Dion is a PERSON	

Including *Johnny Cash*, there are 21 singer instances found in WordNet, all of them are also included in Wikipedia. The Wikipedia entries for these and similar NEs (e.g., *Celine Dion*) contains information (e.g., "singer", "song-writer"), which is indicative of the relevant semantic categories (e.g., PERSON). See Examples (1) and (2), for more details. Additional information in the form of categories (e.g., American composers) and dominating pronoun type (e.g., personal pronouns, such as "his" and "he") are all very indicative of the semantic class (e.g., PERSON).

Example 1. Johnny Cash (born J. R. Cash; February 26, 1932 – September 12, 2003) was an American singer-songwriter and one of the most influential musicians of the 20th century. Primarily a country music artist, his songs and sound spanned many other genres including rockabilly and rock and roll (especially early in his career), as well as blues, folk and gospel. ... **Categories**: American autobiographers | American Protestants | American Christians | American composers ... 1932 births | 2003 deaths. (based on Freebase Wikipedia Extraction (WEX) released on 2009-0316).

Example 2. Céline Marie Claudette Dion (Fr-Celine-Dion.ogg /se□l□n d□□n/ (help·info)), CC, OQ (born March 30, 1968) is a Canadian singer-songwriter and actress. Born to a large, impoverished family in Charlemagne, Quebec, Dion emerged as a teen star in the French-speaking world after her manager and future husband René Angélil mortgaged his home to finance her first record.[3] In 1990, she released the anglophone album Unison, establishing herself as a viable pop artist in North America and other English-speaking areas of the world.[4] ... **Categories**: Celine Dion | Canadian singers | Canadian female singers | French Quebecers ... 1968 births | Living people. (based on Freebase Wikipedia Extraction (WEX) released on 2009-0316).

Intuitively, by using Wikipedia entries (e.g., *Johnny Cash, Bob Dylan*, *Judy Garland*, and *Lena Horne*) listed under a certain WordNet semantic class (e.g., PERSON), we can extract such features and train a classifier capable of predicting that *Celine Dion* is a PERSON, or more specifically a SINGER.

We present a novel system, *WikiSense*, that automatically learns to classify titles in an encyclopedia (e.g., Wikipedia) into broad semantic categories in an ontology (e.g., WordNet). An example of *WikiSense* classifying the entry of *Celine Dion* in Wikipedia is shown in Table 1. *WikiSense* has determined the indicative features for the entry and goes on to classify *Celine Dion* as a PERSON. *WikiSense* automatically learns the relationships between features and

semantic categories by using WordNet and Wikipedia named entities with minimally supervision. We describe the training process of *WikiSense* in more detail in Section 4.

In our prototype, *WikiSense* return a set of Wikipedia NE titles annotated with WordNet supersenses. By combining *WikiSense* and the traditional handcrafted ontology, WordNet, we can provide a very large scale gazetteer, a potentially useful resource for many NLP tasks.

The rest of the paper is organized as follows. In the next section, we describe Wikipedia and WordNet, the two main resources used in this paper. In Section 3, we survey previous work. In Section 4, We explain in detail the problem statement and proposed methods. Finally in Section 5, we report the evaluation results and error analysis, and conclude in Section 6.

2 Background

2.1 Wikipedia

Wikipedia is a free online encyclopedia compiled by millions of volunteers all around the world. Anyone on the internet can freely edit existing entries and/or create new entries to add to Wikipedia. Owing to the size of its participants, Wikipedia has achieved both high quantity and quality. In fact, the English Wikipedia currently consists of over 2,990,000 articles as of August 12, 2009, and is considered to have similar quality comparing to traditional encyclopedias compiled by experts. (Giles, 2005) Due to these reasons, Wikipedia has become one of the largest referenced tools.

Following the editorial and policy guidelines, the *Manual of Style* (http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style), most Wikipedia articles have homogeneous text structure, metadata, and other consistent characteristics. For example, there is always a "lead section" before the table of content and first heading. The lead section, according to the *Manual of Style*, gives the definition of the title and overviews of the article.

Wikipedia articles also contain rich, handcrafted, structured metadata. These metadata includes information boxes, hand labeled categories, and the template that the article is based on. In Wikipedia, a well developed article may be tagged with up to a few dozens of categories, while articles in the early stage of development are usually tagged with a couple of categories, or even no category at all. The entire Wikipedia category structure consists of more than 350,000 categories, some appear in many articles, while many appear in only a handful of articles. These categories are a mixed bag of subject areas, attributes, hypernyms, and editorial notes.

Internal linking, a phrase in one entry linking to another entry in Wikipedia, is also one of the most important and unique features of Wikipedia setting it apart from traditional offline encyclopedias. The article body usually contains dense internal links annotated by volunteer editors. These internal links provide the readers with a simple way to access important relevant topics in Wikipedia. Following the guidelines in the *Manual of Style*, only important and closely related keywords are selected and linked.

In Wikipedia, every articles has a unique and case sensitive title, which is also used as in the last part of its URL. When title conflict arises, a parenthetic explanation is appended to the conflicting titles rendering them unambiguous, and that will also trigger the creation of a *disambiguous page* (see http://en.wikipedia.org/wiki/Wikipedia:Disambiguation) that lists links to each of the conflicting titles with a short description. For example, the term *Saturn* can refer to many articles with a different meaning, including a planet, a Roman god, a magazine, and a microprocessor. Each of these articles has its own unique title, such as *Saturn*, *Saturn* (*mythology*), *Saturn* (*magazine*), and *Saturn* (*microprocessor*). All of these resolved titles are listed in the disambiguation page entitled *Saturn* (*disambiguation*).

2.2 WordNet

WordNet is a lexical semantic database for the content words in English, initially to support psycho-linguistic research, but increasingly being used in natural language processing, information retrieval, and artificial intelligence research and applications. The development of

WordNet began in 1985 at Princeton University, and many freely-downloadable versions of the database have been released under a BSD style license. WordNet groups nouns, verbs, adjectives, and adverbs into sets of synonyms called synsets, and provides glosses with definitions and examples. Every synset contains a group of synonymous words or phrases with different senses of a word are in different synsets. For ambiguous words, WordNet provides indication of how often a word appears in a specific sense (synset) with the estimated frequency count based on a semantically tagged corpus (i.e., SEMCOR).

Unlike other dictionaries, WordNet records various semantic relations between synsets. In the latest Version 3.0, the database contains extensively a total of 207,000 word-sense semantic relations between 150,000 words organized in over 115,000 synsets. For example, the synset of {*good, right, ripe*} is defined as *most suitable or right for a particular purpose* and exemplified with three sentences: *a good time to plant tomatoes*; *the right time to act*; *the time is ripe for great sociological changes*). The semantic relations connecting synsets include (1) nominal relations: *hypernyms, hyponyms, coordinate terms, holonym, meronym, hypernym for nouns* (2) verbal relations: *hypernym, troponym, entailment, coordinate terms* (3) adjectival relations: *related nouns, similar to, participle of verb*, and (4) adverbial relations: *root adjectives*. There are also relations connecting a specific word/phrase in a synset, relations including *antonyms* and *morphological derivatives*. These relations are extensively exploited in many word sense disambiguation study and information retrieval research.

With the *hypernymy* relations, one can derive a hierarchical semantic classification of entities or actions. Additionally, WordNet also allocates the synsets into lexicographer files, also known as lexical files or *supersenses*, for division of work among lexicographers. With lexical file information, one can derive a flat semantic classification, which in some situation is more convenient to use. However, it is not always easy to classify a named entity into one of the 26 nominal lexicographer files because of the systematic ambiguity between LOCATION and OBJECT (e.g., island), GROUP and ARTIFACT (e.g., bank), ANIMAL and FOOD (e.g., fish), just to name a few. In this paper, we propose to classify Wikipedia named entities into WordNet lexicographer files categories. Since NEs concentrate in 9 classes, we ignore other 17 classes for simplicity. Supersense tagging of Wikipedia title is a preliminary step for providing more specific semantic linkage between Wikipedia and WordNet.

3 Related Work

Mining semantic knowledge from Wikipedia is an increasingly active area of research. Wikipedia is becoming increasingly popular and is considered a useful resource in the field of natural language processing. Wikipedia represents a gold mine of information, attracting a growing community of researchers to tap into this huge, constantly evolving repository of concepts and relations in order to apply to many interesting tasks. Recently, works involve mining meaning from Wikipedia has been summarized in Medelyan et al. (2009), a survey paper that provides an in-depth description of the creation process and textual structure of Wikipedia pages, and review diverse effort that exploits Wikipedia in many research areas: word sense disambiguation (Mihalcea and Csomai, 2007; Ruiz-Casado et al., 2005; Wu and Weld, 2007), bilingual lexicography (Erdmann et al., 2008; Tyers and Pienaar, 2008), multilingual information retrieval (Potthast et al., 2008), information extraction and question answering (Banko et al., 2007; Toral and Muñoz, 2007; Ferrández et al., 2007), named entities recognition (Bunescu and Paşca, 2006; Cimiano and Volker, 2005; Dakka and Cucerzan, 2007), and ontology construction (Auer et al. 2007; Ponzetto and Strube, 2007; Suchanek et al., 2007).

In our work we address a specific knowledge mining problem of identifying and categorizing named entities that appear as the titles in Wikipedia (e.g., identifying and categorizing *John Updike* and *Terrorist* as named entities of the types PERSON and COMMUNICATION).

In the work on mining meaning from Wikipedia, the most closely related body of research to our work focuses on word sense disambiguation of Wikipedia titles, words and phrases.

Mihalcea (2007) considers the same title (e.g., *bar*) with different parenthetical explanations (e.g., *bar* (*establishment*) and *bar* (*counter*)) as sense inventory. Using Wikipedia titles as sense inventory has the advantage of having available the running text with hyperlinks to these titles as the training data for WSD. More recently, Mihalcea and Csomai (2007) introduce the *Wikify!* system for identifying keywords and hyper-linking them to relevant Wikipedia entries, based on automatic keyword extraction and word sense disambiguation techniques. Their proposed methods rely on two characteristics of Wikipedia, the parenthetical explanations (for disambiguation) and the internal link structure. Similarly, we use high-frequency parenthetical explanations to generate titles/entries for certain semantic classes in WordNet as the training data for broad sense classification of titles (with or without a parenthetical explanation).

In effort to utilize the information provided in Wikipedia categories, Suchanek et al. (2007) developed the YAGO ontology which contains links from Wikipedia categories to WordNet senses, thereby resolving the ambiguities that exist in category terms (e.g., *Capitals in Asia* is related to *capital city*, while *Venture Capital* is related to *fund*). Although YAGO only covered about 50% of all Wikipedia categories, these categories include substantial part of Wikipedia articles (over 90%).

In a study more closely related to our work, Ciaramita and Johnson (2003) also proposed a method for automatically expanding WordNet by supersense tagging out of vocabulary words/phrases in a corpus. In contrast, we focus on supersense tagging Wikipedia titles. By exploiting characteristics in Wikipedia, our approach is able to achieve high accuracy.

4 Proposed Method

In this section, we formally state the problem we are addressing, and explain in detail the proposed methods for automatically generating training data and feature extraction.

4.1 Problem Statement

Given a lexical ontology, in which its vocabulary consists mainly common nouns and a small portion of named entities. Our goal is to automatically expand the vocabulary of the given ontology (e.g., WordNet) to cover more named entities. For this, we use the given ontology to provide semantic categories for classifying OOV. Other resources used in our proposed method include Wikipedia and YAGO. In this work, we focus on assigning only a fixed set of supersenses, which typically are related to named entities in Wikipedia title.

1. person	6. cognition	11. feeling	16. event	21. time
2. state	7. possession	12. attribute	17. quantity	22. shape
3. body	8. phenomenon	13. relation	18. motive	23. plant
4. act	9. substance	14. process	19. animal	24. object
5. food	10.communication	15. location	20. artifact	25. group

Figure 1: List of lexicographer files for NEs (shaded), not including noun.Tops.

In Figure 1, we show the 25 nominal lexicographer files in WordNet, which we called supersenses, excluding noun.Tops. Since we are focusing on classifying NEs, naturally not all 25 supersenses are targeted. Our proposed method concentrates on 9 supersenses that are typically related to NEs and ignore other supersenses (e.g., SUBSTANCE) not often related to NEs. For example, the SHAPE category typically contains common nouns.

4.2 Minimal-Supervision Training

In the first stage of training, we identify and retrieve all Wikipedia entries with an NE title. We perform this identification task simply by checking the title's upper/lower letter instances appearing in the article body. In an implementation of the proposed method and for a recent Wikipedia dump, we retrieved 1,736,645 articles (out of 2,307,815) with an NE title.

In the second training stage, we automatically generate the training data for semantic classification from the retrieved Wikipedia NE entries and exploit several different characteristics of Wikipedia to extract features for constructing a maximum entropy classification model. Feature extraction is explained in detail in the following subsections. Finally, supersense tagging is performed to all Wikipedia entries with an NE title, and a supersense ontology of NEs is created.

4.2.1 Automatically Generating Training Data

We propose two approaches for generating training data, one fully automatic and one requires minimal hand-labeling. These two methods can be used independently, therefore, our method becomes unsupervised if only the training data generated by the first approach is used. However, experimental results show that combining both approaches generates a larger training set and therefore the model trained on this larger dataset outperforms the model trained on the dataset created fully automatically.

Our first approach collects Wikipedia article with an NE title listed in WordNet. These titles are then associated with the corresponding supersense in WordNet and are used as the training data. To produce high quality training data, we exclude all ambiguous candidates. These excluded candidates either have a parenthetical explanation in title, indicating a conflicting title in Wikipedia (ambiguous in Wikipedia), or correspond to more than one sense in WordNet (ambiguous in WordNet). This method produces 5,411 training entries with high quality.

The second approach is based on the parenthetical explanation appended to conflicting titles used to resolve ambiguous titles. Some these parenthetical explanations are very indicative of the supersense of title, while others are more of a general topic. For example, the parenthetic explanation of "*(album)"* is related COMMUNICATION named entities, while the parenthetic explanation of "*(United State)*" is related to many things, including *Social Security*, *Electoral College*, and *Democratic Party*, and is not necessarily indicative of any specific supersense. Using this property, we handpicked and hand-labeled 30 parenthetical explanations with supersenses, generating 42,645 Wikipedia entries that can be used as the training data .

Figure 2 shows the 30 hand-labeled parenthetical explanations, selected from top-rank parenthetical explanations, for generating a training data set that complements the training data generated in the first approach.

COMMUNICATION	album, film, song, TV series, novel, magazine, soundtrack, EP, play, musical, software, opera, news paper, video game	GROUP	band, rugby league, company, rugby union, school, company
LOCATION	UK Parliament constituency	OBJECT	crater (e.g., *Wood (crater*))
ARTIFACT	Amtrak station, automobile	PERSON	footballer, politician, actor
LOCATION	village (e.g., *Duty (village)*)	ANIMAL	horse (e.g., *Sweep (horse)*)

Figure 2: The 30 hand-labeled parenthetic explanations.

4.2.2 Training Features from Article Content

The body of an Wikipedia article contains rich information helpful for semantic classification of title. Although free text are more difficult to process comparing to well structured metadata, the structure of the articles in Wikipedia is in some degree homogeneous. Therefore, by exploiting this property, we can effectively extract relevant information from the article body as features. In this work, only the lead section of Wikipedia articles is used based on the observation that this section usually is the abstract that contains key information.

Characteristic of all Wikipedia article, the first sentence of the lead section usually gives a simple definition to the title. In many cases, we can see that the first sentence of the leading section describes the genus of the title word/phrase. For example, the first sentence of the entry *Blog* is "A blog is a type of website, usually maintained by an individual ...", indicating the title "Blog" refers to a type of *website* (the genus). Using WordNet, we can determine "website" belonging to the lexicographer file COMMUNICATION. Since the semantic class of the title

word/phrase is relevant to the genus, we use genus and its supersense as important features. Table 2 shows two example articles in Wikipedia and their feature, including *Michael Jackson* and *France*.

Table 2: Example of genus extraction from the first sentence of three entries.

Title	Defining Sentence	Genus	Supersense
Michael Jackson	Michael Joseph Jackson was an American recording artist, entertainer ...	artist, entertainer	PERSON, PERSON
France	France is a country located in Western Europe, ...	country, Europe	LOCATION, LOCATION

Pronouns appearing in Wikipedia article body often reveal the semantic category of the title word/phrase. Pronoun resolution is still an open problem in natural language processing, and resolving pronouns in a large corpora like Wikipedia is infeasible. As a characteristic of all encyclopedias, every entry in Wikipedia has a central topic, and does not contain much off-topic information, especially in the lead section. Therefore, the dominating pronoun type is likely to refer to the title. Therefore, instead of resolve all pronouns in the article to find out which type of pronoun refers to the title, we took an alternative way of using a statistic method which simply counts redundancy and assume that the dominating pronouns (used over 50%) as referring to the title.

We group pronouns into the following three groups according to the type of referents: PERSON (e.g., he, she, him), OBJECT (e.g., it, its, itself), and GROUP (e.g., they, them, themselves). For example, in the lead section of Wikipedia entry *William Shakespeare*, there are 24 instances of pronouns in the PERSON class, one OBJECT pronoun, and no GROUP pronouns, suggesting that the topic is a PERSON.

4.2.3 Features from Metadata

Metadata also contains rich information helpful for determining the semantic category of the title. A simple approach is to simply use the collaboratively labeled categories as features. As previously mentioned, these categories are a mix bag of subject areas, attributes, hypernyms, and editorial notes, therefore may contain noise and ambiguity. Moreover, most of these categories are very specific, even more fine-grained than WordNet word senses. For example, instead of just "*writer"*, the *Charles Dickens* page is tagged with *English short story writer*, *English historical novelist*, and *Literature collaborators*. It is clear that we should cluster these closely related categories to a single WordNet sense, writer#n#1 or the supersense PERSON, to help providing more general feature.

Wiki Title	Paul Jorion
Categories	Consciousness researchers and theorists, Artificial intelligence researchers, Belgian writers, Belgian sociologists, Belgian academics
WNSenses	research_worker#1 (2), writer#1 (1), sociologist#1 (1), academician#3 (1)
Supersense	noun.person (5)

Wiki Title	Miaoli City
Categories	Cities in Taiwan, Taiwan geography stubs(not included as features)
WNSenses	city#1 (1)
Supersense	noun.location (1)

Figure 3: Example of features generated from categories.

For this in mind, we used YAGO to map the overly specific Wikipedia categories to more general WordNet senses. First, categories are mapped to WordNet senses via YAGO as features (lower level of clustering). To generate the second group of features, WordNet is used to transform the mapped senses to its supersenses (higher level of clustering). To further increase coverage for entries not covered by YAGO, we also used the original categories as the third group of features (no clustering). Figure 3 shows the features and their redundancy extracted from two articles in Wikipedia. Notice that the second article, entitled *Miaoli City*, contains the category of *Taiwan geography stubs* which is an editorial note indicating that the entry is still in the initial stage of development. Editorial categories indicating stub articles are not included as our training features. However, even underdeveloped entries may still contain relevant information for classifying the title, we do not completely exclude these articles at training time.

4.3 Runtime

Once the classifier is trained, supersense tagging is performed to the entire corpus using the same feature extraction method. Thousands of new and unfinished articles are created by volunteers or robots daily. Initially receiving a stub status, these new articles are less accurate, contain less information, and are tagged with a couple of or no categories. Moreover, entries in Wikipedia can virtually be created by anyone on the Web. New entries are sometimes created by new volunteers less familiar with Wikipedia than experienced editors. An underdeveloped entry may sometimes be removed due to policy issue or merged to an existing and well developed entry that has an overlapping topic. Therefore, unlike in the training stage, we exclude all stub or under-tagged articles at runtime for higher accuracy. In spite of excluding these articles, new entries that are popular and/or important are likely to develop quickly. Experimental results show excluding all stub articles, articles that generate less than four features, and classification results with lower probability still achieves a high applicability of 78.8%, covering 1,255,532 entries, about six times the scale of WordNet in terms of vocabulary size.

5 Evaluation

5.1 Experimental Setting

First, we describe our experimental setting and statistics for the following experiments. Our source for Wikipedia dump is the Freebase Wikipedia Extraction (WEX) released on May 16, 2009, which we then exclude all pages irrelevant to our work (disambiguation pages, 'list of' pages, and other special pages). Tsujii's POS tagger (Tsuruoka and Tsujii, 2005) was applied to the first sentence of every article to extract genus terms. WordNet 3.0 is used to generate Training Set #1, and the 30 hand-picked and hand-tagged parenthetic explanations in Figure 2 are used to generate Training Set #2. Table 3 shows the number of entries in the Wikipedia corpus ready for WikiSense processing, the training sets, and the produced semantically classified named entities.

In the first experiment, we show compared results of the unsupervised method (using only Training Set #1) and minimally supervised method (Training Set #1 and #2). We then analyze the importance of different types of features by showing experimental results of removing one type of features at a time and by using only one type of features at a time.

Table 3: Statistics for experimental settings and results.

Corpus	# of entries
Wikipedia	2,307,815
NE Titled Entries	1,736,645

Training Data	# of entries
Set #1	5,411
Set #2	42,645

Result	# of terms
WikiSense	1,255,532

5.2 Experimental Results and Analysis

Table 4 shows comparison of using only Training Set #1 (unsupervised) and using both training sets (minimally supervised). We can see that using only training data generated by WordNet is

not satisfactory, mainly due to the uneven coverage of WordNet on NEs. For example, 337 {*author, writer*} instances are included in WordNet, but only three book instances are included (*Das Kapital*, *Erewhon*, and *Utopia*). There are also no magazines, movies, or music records. Results show by simply tagging 30 parenthetic explanations, we effectively expanded the coverage of the training data, and significantly improve the precision rate.

Table 4: Comparison of using different methods to generate training data.

Method	Precision	Applicability
Unsupervised	64.4%	78.6%
Minimally Supervised	95.4%	78.8%

To evaluate the impact of removing each type of features on precision and applicability, we remove the threshold on number of features. Results in Table 5 (a) show removing features generated from Wikipedia categories has the highest impact on both applicability and precision, removing genus features only affected precision, and removing pronoun features has little impact.

Table 5: Evaluation of feature importance by feature removal (a) and single feature (b).

	Precision %	Applicability %
All Features	95.4	78.8
Remove Category	77.6 (-17.8)	70.6 (-8.2)
Remove Genus	86.5 (- 8.9)	78.8 (-0.0)
Remove Pronoun	95.1 (- 0.3)	78.8 (-0.0)

(a)

	Precision	Applicability
All Features	95.4	78.8
Category Only	85.8	78.8
Genus Only	76.2	64.6
Pronoun Only	75.3	29.2

(b)

In our last experiment, we removed the threshold on number of features, and use each type of features alone to evaluate the contribution of each type of features. Results in Table 5 (b) show features generated from Wikipedia categories are most crucial both for precision and applicability. Features that uses dominating pronouns is only applicable on 29.2% of the entries, but still achieved 75.3% precision.

6 Conclusion

In summary, we have presented a novel method for learning to classify named entities into broad semantic categories. The method involves generating training data by labeling high-frequency parenthetical explanations, extracting lexical, conceptual, and metadata features to train a statistical classifier. We have implemented and thoroughly evaluated the method as applied to Wikipedia and WordNet. The experimental results show that the method effectively exploit the rich, consistent, and indicative features in Wikipedia pages to classify titles and extend WordNet's coverage on named entities.

Many avenues present themselves for future research and improvement of our system. For example, existing methods for pronoun resolution could be implemented to improve the quality of pronominal feature. Natural language parsing techniques could be used to identify genus terms more precisely. Additional parenthetical explanations could be labeled to increase the amount of training data and support more semantic categories. Effort could be made to balance the distribution of the training data among semantic categories. Additionally, an interesting direction to explore is classifying titles into more specific categories, thereby linking them to direct hypernyms in WordNet and establishing "has instance" relations. For instance, it would be useful for some applications to classify *Celine Dion* as a SINGER. Yet another direction of research would be classifying OOV common nouns, especially in a specific domain. For example, it would be interesting to classify terminology (e.g., the environmental protection term, *ecological footprint* in Wikipedia) broadly as an ATTRIBUTE, or perhaps even more specifically as the direct hypernym of WordNet word sense (i.e., *footprint#n#3*).

References

Auer, S., C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak and Z. Ives. 2007. DBpedia: a nucleus for a web of open data. *Proceedings of the Sixth International Semantic Web Conference and Second Asian Semantic Web Conference*. vol. 4825, pp. 715–728.

Banko M., M.J. Cafarella, S. Soderland, M. Broadhead and O. Etzioni. 2007. Open information extraction from the Web, *Proceedings of the 20th International Joint Conference on Artificial Intelligence.* pp.2670-2676.

Bunescu, B. and M. Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics.* pp.9-16.

Cimiano, P. and J. Volker. 2005. Towards large-scale, open-domain and ontology-based named entity classification. *Proceedings of the Internatioal Conference on Recent Advances in Natural Language Processing*. pp.166-172.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pp.708-716.

Dakka, W. and S. Cucerzan. 2008. Augmenting Wikipedia with Named Entity Tags. *Proceedings of the Third International Joint Conference on Natural Language Processing.*

Erdmann, M., K. Nakayama, T. Hara and S. Nishio. 2008. An approach for extracting bilingual terminology from Wikipedia. *Proceedings of the 13th International Conference on Database Systems for Advanced Applications.*

Fellbaum, C. 1998. WordNet an Electronic Lexical Database. *MIT Press, Cambridge, MA.*

Ferrández, F., A. Toral, Ó. Ferrández, A. Ferrández, and R. Muñoz. 2007. Applying Wikipedia multilingual knowledge to cross-lingual question answering. *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems*. pp.352-363.

Giles, J. 2005. Internet encyclopaedias go head to head, *Nature* 138 (15).

Medelyan, O., D. Milne, C. Legg and I.H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies 67.* pp.716-754.

Mihalcea, R. and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge, in *Proceedings of the sixteenth ACM conference on information and knowledge management*. pp.233-242.

Ponzetto, S.P. and M. Strube. 2007. Deriving a large scale taxonomy from Wikipedia *Proceedings of AAAI.* pp.1440-1445.

Potthast, M., D. Stein, and M.A. Anderka. 2008. Wikipedia-based multilingual retrieval model. P*roceedings of the 30th European Conference on IR Research*.

Ruiz-Casado, M., E. Alfonseca and P. Castells. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. *Proceedings of AWIC.*

Suchanek, F.M., G. Kasneci and G. Weikum. 2007. Yago: a core of semantic knowledge. *Proceedings of the 16th World Wide Web Conference, WWW*.

Toral, A. and R. Muñoz. 2007. Towards a named entity WordNet (NEWN). *Proceedings of the Sixth International Conference on Recent Advances in Natural Language Processing*. pp.604-608.

Tsuruoka, Y. and J. Tsujii. 2005. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data, *Proceedings of HLT/EMNLP*, pp.467-474.

Tyers, F. and J. Pienaar. 2008. Extracting bilingual word pairs from Wikipedia. *Proceedings of the SALTMIL Workshop at Language Resources and Evaluation Conference.*

Wu, F. and D. Weld. 2007. Autonomously semantifying Wikipedia. *Proceedings of the 16th ACM Conference on Information and Knowledge Management*. pp.41-50.