Classifying Taxonomic Relations between Pairs of Wikipedia Articles

Or Biran
Columbia University
Department of Computer Science
orb@cs.columbia.edu

Kathleen McKeown
Columbia University
Department of Computer Science
kathy@cs.columbia.edu

Abstract

Natural language generation systems rely on taxonomic thesauri for tasks such as lexical choice and aggregation. WordNet is one such taxonomy, but it is limited in size. Motivated by the needs of a generation system in the scientific literature domain, we present a method for building a taxonomic thesaurus from Wikipedia articles, where each article represents a potential concept in the taxonomy. We propose framing the problem of creating a taxonomy as a classification task of the potential relations between individual Wikipedia article pairs, and show that a supervised algorithm can achieve high precision in this task with very little training data.

1 Introduction

Thesauri are useful resources for many NLP applications. In particular, taxonomic thesauri which contain synonymy and hypernymy relations are important for natural language generation (NLG) systems which must make decisions regarding lexical choice and aggregation. WordNet (Fellbaum, 1998) is one such thesaurus which has many uses in generation (Jing, 1998), but its set of concepts (called *synsets*) is quite limited. It does not contain many domain-specific concepts, nor does it contain technical concepts that emerged very recently. This work is motivated by the needs of a NLG system in the scientific literature domain, where these missing concepts are absolutely necessary for any practical application. Our goal is to generate a thesaurus containing synonomy and hypernomy relations between scientific terms which a generation system can use to select the most appropriate term given a context.

The English Wikipedia has over 4 million articles, and over 8.6 million titles if *redirects*, which are alternative titles for the articles, are included. These titles are essentially lexical terms referring to concepts. Crucially, it contains articles describing many domain-specific concepts, and, in particular, scientific and technological concepts. For example, Wikipedia contains articles with titles such as *Supersymmetric String Theory*, *Gorilla Glass* and *Sentiment Analysis*, all of which are missing from WordNet. While there have been attempts to build ontologies from Wikipedia, these tended to focus (in their optimization and evaluation) on entities such as people, places and events. There is still a need for a WordNet-like taxonomy which would contain accurate synonymy and hypernomy relations for highly specialized terms from various scientific domains (for our purposes) and other specialized domains.

Unlike previous approaches, which tend to rely on WordNet's hierarchy and/or on Wikipedia's pseudo-hierarchy of *categories*, we frame the problem as a binary classification task for a pair of Wikipedia article titles - deciding whether the term representing the concept in the first article is a hypernym of the term representing the second or not. This enables us to handle specialized concepts which are far from the established concepts in the WordNet hierarchy.

WordNet-like taxonomies behave in some ways as a dictionary, in others as an ontology. To avoid confusion, we define the main terms we use in this paper and what they correspond to:

- A *concept* in computational ontologies is a unique semantic entity. We assume that WordNet synsets correspond to concepts. Another assumption we make is that each Wikipedia article describes something analogous to a concept; this assumption does not work for some types of articles (e.g. Template articles), and we remove such articles before processing, as explained in section 3.

International Joint Conference on Natural Language Processing, pages 788–794,
Nagoya, Japan, 14-18 October 2013.

- A *term* is a lexical entity (word or combination of words) used to refer to a concept. Each WordNet synset contains multiple terms (synonyms) which all refer to the concept represented by the synset. We treat Wikipedia article titles as terms referring to the concept described in the article. In addition to the main title, Wikipedia has multiple additional *redirect titles* referring to each article. We do not *a priori* treat these as synonyms, as they are often hypernyms, hyponyms or even terms referring to distinct (though related) concepts (for example, at the time of this publication, *Disambiguation* redirects to *Word Sense Disambiguation*; *nano-SIM* redirects to *Subscriber Identity Module* (SIM); and *Sheep Sounds* redirects to *Sheep*).
- *Relations* in this work are semantic relations between pairs of terms - specifically, synonymy and hypernymy. This is in contract to the use of the word in ontologies where relations occur between pairs of concepts.

The following are a few examples of relations that do not appear in WordNet and which our method correctly finds:

- *Gene Silencing* is a hypernym of *RNA Interference*
- *Graph Property* is a hypernym of *Clustering Coefficient*
- *Conditional Random Field* and *CRF* are synonyms

We will use these examples to illustrate the limitations of other methods in the next section.

2 Related Work

There have been many attempts to extend WordNet with concepts from Wikipedia. Because WordNet has some of the properties of an ontology, most work on extending WordNet with Wikipedia concepts was in the context of creating an ontology. Although our work is different in that we focus on extending only the taxonomic relations between the terms, this related work is still very relevant. There have also been attempts to create ontologies directly from Wikipedia in various ways, and we discuss those as well.

Yago (Suchanek et al., 2007) is a large ontology (over 10 million concepts) based on WordNet and extended with concepts from Wikipedia and other resources. Its hypernymy hierarchy (a relation called *subClassOf*) is derived by matching articles with existing WordNet synsets using the lexical and syntactic properties of the title. This approach works well for some complex entities: a title like "American people in Japan" contains the head compound *people* which matches the WordNet synset *Person/Human*. It does not work as well for scientific concepts, where titles tend to be less clearly related. For example, Yago contains the concepts *Clustering Coefficient* and *RNA Interference*, because they are titles of Wikipedia articles; but these concepts are not part of the *subClassOf* hierarchy, because their titles are not lexically similar to *Graph Property* and *Gene Silencing*, repectively.

Ponzetto and Navigli (2009) link Wikipedia *categories* to existing WordNet synsets, leveraging the category structure to enrich WordNet with concepts from Wikipedia. Wikipedia categories are mostly thematic, with no strict hierarchical structure and do not represent a taxonomy, but they do tend to be somewhat hierarchical for concepts low in the hierarchy (i.e., more specific concepts). For example, *Public transport in Stockholm* is in the category *Public transport in Sweden* which is in the category *Public transport*, and the latter corresponds to a synset in WordNet. However, this is not true for many scientific concepts, where even the more general concept does not appear in WordNet. For example, *Clustering coefficient* is in the category *Graph invariants*, but the categories above that are purely thematic, and WordNet does not contain a synset for *Graph invariant*. Similarly, the term *CRF* is the title of a disambiguation page, which does not belong to any categories and so would not be linked to *Conditional Random Field*.

Syed and Finin (2010) match each Wikipedia article to a WordNet synset as a hypernym-like superclass. Their method relies on the synset-category mappings of (Ponzetto and Navigli, 2009), extending it with information obtained from the hyperlink structure of the Wikipedia articles. However, this approach is still limited by the choice of categories for each article. In addition, it does not work as well for articles with a small number of hyperlinks, which is typical of the more specialized scientific articles.

There have also been attempts (Auer et al.,

2007; Wu and Weld, 2008) to build ontologies from the *infoboxes* of Wikipedia articles, which commonly occur in articles of (e.g.) people and places but not in the articles of most domain-specific concepts.

There has also been work mapping words from Wikipedia articles to particular senses within WordNet using WSD techniques (Mihalcea, 2007; Milne and Witten, 2008). Our work is different in that we attempt to create a thesaurus specifically containing terms that are not in WordNet.

2.1 Contrast to Related Work

In addition to not being optimal for the scientific domain, these approaches all have in common that in attempting to extend WordNet using Wikipedia they rely on the structural information in WordNet directly. This generally means that the further down the hierarchy a term is (that is, the further it gets from the most specific hypernym available in WordNet) the less accurate the constructed taxonomy becomes with regard to its relations. This again works well for some entities, where WordNet contains reasonably specific concepts (e.g., occupations and nationalities for people, industries for organizations) but not too well for specialized concepts in specific domains.

In contrast, in our approach, WordNet is only used to provide the labels for very few relations ($5,000$) that are used in training and (separately) in evaluation. However, these relations are all considered individually. We do not rely on the WordNet hierarchical structure as a whole; instead, we learn to classify the relation between a pair of terms using only information from their Wikipedia article content. This makes our method more robust with regard to very specific concepts. Evaluating other methods using gold data from WordNet may be biased, because concepts from WordNet (even if they are not used directly in ontology construction) are inevitably close to other concepts in WordNet. It can be expected that for more highly specialized concepts, these methods will not perform as well. In our approach, there is nothing special about a relation whose concepts appear in WordNet, and performance on those should give a good indication of performance on other relations (perhaps with the caveat that concepts which appear in WordNet may have larger corresponding articles on average).

3 Data and Definitions

Since we want our terms from Wikipedia to refer to concepts, we remove from the Wikipedia corpus all the pages whose title begins with a wikipedia special prefix. These prefixes are single words followed by a colon, and denote a special type of wikipedia page, such as Template, Category or File. We also remove all pages whose title does not contain at least one English letter character.

We define a Wikipedia term as any Wikipedia article title and any redirect title which passes the filters above. This lexical definition is motivated by the need to find synonymy and hypernymy. It also makes evaluation (which we do using WordNet) more straightforward. To make things even simpler, we completely ignore senses. While word sense disambiguation has been a major part of some related work, it is less crucial for our purposes since specialized terms are less likely to be ambiguous than general terms. We hypothesize that the Wikipedia article itself describes the concept that is referred to by the term.

We define a WordNet term as any term (synonym) participating in any noun synset in WordNet. Wikipedia terms are matched to WordNet terms lexically, with some pre-processing: we lowercase the titles, replace underscores with spaces, remove diacritics from unicode characters and remove text in parentheses (which are commonly used in Wikipedia to disambiguate senses).

Using our definition, there are 117,092 WordNet terms. The total number of potential terms from Wikipedia is 9,096,022, which covers 73.62% of the WordNet terms. WordNet has 494,892 hypernym and synonym relations between all terms. The set of all potential relations from the Wikipedia term set (which is $9,096,022^2$ in size) covers 63.71% of those.

We define our task as a binary classification over all potential relations from the Wikipedia term set. For each ordered pair of terms, we want to decide whether the first is a hypernym of the second or not. If two terms are determined to both be hypernyms of each other we treat them as synonyms. We evaluate on a dataset sampled from that subset of the Wikipedia terms which also exist in WordNet.

To determine the relations for all Wikipedia terms, the space of potential relations must first be dramatically reduced from its current size of over 82 trillion data points. In this paper, we present

results on sampled subsets.

4 Features

We extract fourteen features of four general types. For most of these, it is essential that each term in the pair corresponds to a Wikipedia article. Each term matches either the article title, or a redirect title that redirects to the article.

4.1 Features from the hyperlink structure of Wikipedia

We utilize the graph structure of hyperlinks between articles to build the following eight features:

1. First article links to second (yes or no)
2. Second article links to first (yes or no)
3. The cosine similarity between the outgoing links of the articles
4. The ratio of outgoing links in the first article shared by the second article
5. The ratio of outgoing links in the second article shared by the first article
6. The cosine similarity between the incoming links of the articles
7. The ratio of incoming links in the first article shared by the second article
8. The ratio of incoming links in the second article shared by the first article

One of the powerful aspects of Wikipedia is its hyperlink structure. Based on the simple assumption that article A links to article B only if the information in B is related to or somehow assists in understanding the information in A, the intuition is that two articles having a semantic relation will more often link to one another, and will in general link to more similar (additional) articles than will two unrelated articles. The Wikipedia hyperlink structure has been used to compute similarity between articles, for example in (Syed and Finin, 2010) and (Yazdani and Popescu-Belis, 2010).

Wikipedia links contain two bits of information: the title of the article they link to, and the text of the hyperlink as it appears in the referring article. For features (1) and (2), we allow both: that is, even if a hyperlink links to a third article, but uses the relevant article's title in the text,[1] we count that as a link to the relevant article. For the other features, we use only the title of the actual linked articles. The reason is that in features (1) and (2) we want to measure something different than in the rest: whether or not one of the articles mentions the other directly (hyponyms often mention their hypernyms, while hypernyms sometimes list their hyponyms). An article being mentioned by name in a hyperlink, even when the link goes elsewhere, answers that criteria. The other features are intended to capture the similarity of the two articles based on how related the links to/from them are, and so using the text is less relevant (and that information would be captured to some extent by the feature in the next category instead).

[1]For example, a link for the article *New York City* may have only *New York* in the text, which is the title of an article about the state

4.2 Features from the text of the articles

For each article, we build a bag-of-words vector. These vectors are used to compute the cosine similarity between the two articles of a pair, which we use as a feature.

The intuition behind this central feature is that articles having a semantic similarity will also have a higher lexical similarity. This is the same intuition behind distributional similarity (Church and Hanks, 1990), which is that terms surrounded by similar context tend to be semantically related. In this case, the context does not surround the terms but is in the body of the articles corresponding to them. Lexical similarity between Wikipedia articles has been used successfully to link articles, for example in (Yazdani and Popescu-Belis, 2010).

4.3 Features from the redirect structure of Wikipedia

The Wikipedia dump contains a list of redirects from multiple alternative titles to each article. We use those to build three boolean features:

1. The first term redirects to the second term's article (yes or no)
2. The second term redirects to the first term's article (yes or no)
3. Both terms redirect to the same, third article (yes or no)

As mentioned earlier, redirect titles are often synonyms, hypernyms or hyponyms of the main title of the article they redirect to. While it is not consistent enough to use as a strict rule, this structure can be taken advantage of in features.

4.4 Features from the terms (i.e. the article titles)

In some cases, the terms themselves can point at the relation among them. In particular, hyper-

nyms are sometimes lexical subsets of their hyponyms (*String Theory* is a hypernym of *Super String Theory*; *Leukemia* is a hypernym of *lymphocytic leukemia* which in turn is a hypernym of *B-cell chronic lymphocytic leukemia*).

We therefore derive two features from the terms themselves (which correspond to article titles or redirect titles): the difference between the number of words in the two terms, and the number of words which overlap in the two terms.

5 Method and Evaluation

Our training, development and test data sets all consist of ordered pairs of terms from Wikipedia where both terms also appear in WordNet. The label is positive if the first term in the pair is a hypernym (or a synonym) of the second. The positive samples (which consist of pairs exhibiting either hypernymy or synonymy) are sampled from the relations in WordNet. To get negative samples we randomly pair terms from WordNet that have no relation between them.

We train two SVM classifiers: one on a small training set of 5,000 labeled pairs, and the other on a much larger set of 100,000 pairs. In both cases, the training sets are balanced and we used a balanced development set of 186,000 pairs. We then evaluate on a large unbalanced test dataset of 10 million pairs. Using the number of WordNet's total potential relations ($117,092^2$) and the number of its true relations (494,892), we estimate the ratio of real relations in the natural set of all potential relations to be around 0.0036%. Estimating the factor by which we aim to reduce the size of the total space (of 82 trillion) as 1,000, the test set is then built using 360,000 sampled true relations from WordNet, while the rest are randomly paired concepts (which appear in WordNet but have no relation between them).

To illustrate our performance specifically on the science domain, we constructed a second data set using Wikipedia's category hierarchy. In this data set, we included only terms such that their corresponding articles are in a category which is a descendent of the *Science* category with a depth of no more than 20, but are *not* descendents of one of the following categories with a depth of 5 or less: *People*, *Places*, *History*, *Chronology*, *Music*, *Film* and *Sports*. These exclusions are required because descendents of the *Science* category include articles for entities such as scientists and universities, certain historical dates/eras, and expansions of the technologies used in the music, film and sports industries to include entities from these fields (songs, bands, movies...) which then completely overwhelm the data set in size. The depth restrictions are necessary because the category graph is cyclic. In addition to illustrating performance in our intended domain, this test set is important in that it features negative samples that are not entirely random, since they are at least thematically related. The size of this set is 258,971, and it is unbalanced with about 10% positive samples. Note that we use the same classifier (trained on the same unrestricted training set) when evaluating on all test sets, including this one.

To illustrate our approach's advantage over naive methods, we include the results for two baselines. The first uses only the term names and makes predictions based on the Levenshtein distance between them (predicting synonym for distance < 8, hypernym for distance < 12, and none otherwise). The second predicts the relation type based on the lexical cosine similarity between the articles (predicting synonym for similarity > 0.1, hypernym for > 0.05, and none otherwise). The thresholds in both baselines were manually tuned to optimize f-measure on the development set.

In addition, we compare our performance with that provided by querying two leading publicly available ontologies that were constructed using Wikipedia's category hierarchy and infoboxes: Yago (Suchanek et al., 2007) and DBPedia (Auer et al., 2007).

We show two binary evaluations for each data set. The main evaluation, where a positive answer means the (ordered) pair has a hypernymy relation, is shown in Table 1. *SynonymOrNot*, in Table 2, is an additional evaluation over those pairs that were judged as having a relation in the first evaluation, and a positive answer means the pair is a synonym. Recall that we mark as synonyms those pairs that are determined to have both directional hypernyms. We found the results to be statistically significant using a standard t-test.

6 Discussion

The first thing to notice is that the SVM classifiers operate as high-precision, lower-recall systems for both tasks. On the *SynonymOrNot task*, precision is extremely high while retaining a reasonable recall even on the unbalanced test set. This is impor-

	Bal. P	Bal. R	Bal. F	Un. P	Un. R	Un. F	Sci. P	Sci. R	Sci. F
Naive baseline	57.41	69.44	62.85	4.76	69.38	8.91	13.74	80.08	23.45
Lexical baseline	97.14	17.89	30.21	54.31	16.23	24.99	70.22	19.13	30.06
DBPedia	100	0.25	0.5	96.33	0.26	0.52	98.72	1.78	3.5
Yago	100	15.23	26.44	99.96	14.5	25.33	100	29.19	45.19
SVM (trained on 5K samples)	98.75	46.18	62.93	66.03	42.95	**52.05**	64.81	61.23	**62.97**
SVM (trained on 100K samples)	98.13	48.46	**64.88**	57.51	45.22	50.63	57.45	66.3	61.56

Table 1: **P**recision, **R**ecall and **F**-measure obtained for each data set for the main task. **Bal.** stands for **balanced**, the balanced development data set, **Un.** stands for **unbalanced**, the unbalanced test data set, and **Sci.** stands for **science**, the science-only filtered test set.

	Bal. P	Bal. R	Bal. F	Un. P	Un. R	Un. F	Sci. P	Sci. R	Sci. F
Naive baseline	50.76	68.61	58.35	7.02	66.19	12.7	7.65	64.26	13.67
Lexical baseline	68.41	97.83	**80.52**	43.49	97.75	60.2	23.99	92.31	38.08
SVM (trained on 5K samples)	99.92	30.15	46.33	99.65	44.58	**61.6**	97.65	56.12	71.28
SVM (trained on 100K samples)	99.92	29.92	46.05	99.62	44.46	61.48	97.75	58	**72.8**

Table 2: **P**recision, **R**ecall and **F**-measure obtained for each data set for **SynonymOrNot**. **Bal.** stands for **balanced**, the balanced development data set, **Un.** stands for **unbalanced**, the unbalanced test data set, and **Sci.** stands for **science**, the science-only filtered test set.

tant, since a high precision is crucial to maintaining coherence in tasks such as lexical choice.

The classifiers beat both baselines on the main task. The lexical baseline does quite well on the *SynonymOrNot* task, but its performance deteriorates on the unbalanced test sets while the classifiers' performance actually significantly increases due to its high-precision nature.

While the ontologies (Yago and DBPedia) offer incredibly high precision in all cases, their recall is very low (often less than 1% in DBPedia). This is because they focus on entities that are well defined through the category hierarchy and/or infoboxes, which most Wikipedia articles are not.

Overall, the classifiers beat both baselines and both ontologies in both tasks on both test sets. Most importantly, we achieve a relatively high performance on the science domain test set, which is our main goal in this paper.

Finally, it is interesting to note that there is little difference in performance between the SVM when trained on a small training set and when trained on a much larger training set. It seems that whatever can be learned about the data using these features (which is quite a bit, given the performance and especially the precision using this simple approach on a highly unbalanced test set) is learned very quickly, even from a small sampled set.

7 Conclusion and Future Work

We described a simple supervised method of classifying pairs of Wikipedia article titles in terms of the relation among them, covering synonymy and hypernymy. Our approach significantly outperforms the baselines on simulated target data, and achieves very high precision. Unlike previously described approaches, it does not rely on the WordNet hierarchy as a whole, but only on the properties of the individual pair.

In order to use this method in building a taxonomic thesaurus from Wikipedia, we must first reduce the space of potential articles, which is tens of trillions in size. We leave this task and the task of building a full thesaurus to future work. Even without the full thesaurus, our approach can be used to make on-line decisions about the relation between any arbitrary pair of terms.

Acknowledgments

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: a nucleus for a web of open data. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg. Springer-Verlag.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Quang Xuan Do and Dan Roth. 2010. Constraints based taxonomic relation classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1099–1109, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Hongyan Jing. 1998. Usage of wordnet in natural language generation. In *Proceedings of COLING-ACL'98 workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada, August.

Rada Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In Candace L. Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *HLT-NAACL*, pages 196–203. The Association for Computational Linguistics.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *Proceedings of the 21st international jont conference on Artifical intelligence*, IJCAI'09, pages 2083–2088, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Fabian Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In Carey L. Williamson, Mary Ellen Zurko, and Prashant J. Patel-Schneider, Peter F. Shenoy, editors, *16th International World Wide Web Conference (WWW 2007)*, pages 697–706, Banff, Canada. ACM.

Zareen Syed and Tim Finin. 2010. Unsupervised techniques for discovering ontology elements from wikipedia article links. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 78–86, Los Angeles, California, June. Association for Computational Linguistics.

Fei Wu and Daniel S. Weld. 2008. Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 635–644, New York, NY, USA. ACM.

Majid Yazdani and Andrei Popescu-Belis. 2010. A random walk framework to compute textual semantic similarity: a unified model for three benchmark tasks. In *Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC 2010), Carnegie Mellon University, Pittsburgh, PA, USA*, 0.