Measuring Semantic Similarity between Named Entities
by Searching the Web Directory

Jiahui Liu[1] and Larry Birnbaum[2]
Northwestern University
2133 Sheridan Road, Evanston, Illinois 60208 USA
[1]j-liu2@northwestern.edu, [2]birnbaum@cs.northwestern.edu

Abstract

The importance of named entities in information retrieval and knowledge management has recently brought interest in characterizing semantic relationships between entities. In this paper, we propose a method for measuring semantic similarity, an important type of semantic relationship, between entities. The method is based on Google Directory, a search interface to the Open Directory Project. Via the search engine, we can locate the web pages relevant to an entity and automatically create a profile of the entity according to the directory assignments of its web pages, which capture various features of the entity. Using their profiles, the semantic similarity between entities can be measured in different dimensions. We apply the semantic similarity measurement to two knowledge acquisition tasks: thesaurus construction of entities and fine grained categorization of entities. Our experiments demonstrate that the proposed method works effectively in these two tasks.

1. Introduction

Named entities play a vital role in information and language processing. Recently, there has been increasing interest in characterizing semantic relationships between entities. Analogous to the synonym relationship holding between common words, the relationship of semantic similarity holds between entities. Measurement of semantic similarity between entities can provide particular value for tasks concerning the semantics of entities, such as ontology generation, automatic annotation of web pages and question answering.

For common words, general thesauri such as WordNet are important knowledge resources for measuring semantic similarity. However, these general lexical databases provide very limited coverage of named entities. Therefore, similarity measures based on these thesauri are not applicable for named entities.

In this paper, we propose a method for measuring semantic similarity between entities by searching Google Directory [5]. The method utilizes the existing semantic knowledge embedded in the web directory. It does not require any text corpus or domain specific knowledge engineering.

Google Directory uses the links and categories from the Open Directory Project (ODP) [9] and implements its own search functionality on the directory. The Open Directory Project is the most comprehensive human edited hierarchy of web documents. In ODP, web pages are classified into categories in the directory hierarchy according to their topics. Thus, ODP is a rich knowledge resource of human judgment about vast web content. As reported by ODP in September 2006 [9], there are 4,815,303 web pages listed in the directory. Among these web pages, there are many *entity centric pages*, such as commercial web sites of companies, homepages of people, and information pages about locations. When human indexers categorized these pages, they actually categorized them according to the properties of the main entities that the pages were about. Therefore, semantic similarity between named entities can be measured based on the category assignments of their relevant pages in ODP.

In addition, different web pages characterize different aspects of an entity and may be assigned into different categories in the web directory. For example, a person's homepage gives generic information about the person and may be categorized according to the profession of the person; whereas a news web page about the person describes the specific event the person was involved in and may be categorized according to the characteristics of the event. Therefore, using the category assignments of various pages relevant to the entities, we can measure the semantic similarity of the entities in different dimensions.

We demonstrate the value of the proposed method in capturing semantic similarity of entities in two knowledge acquisition tasks. First, we present a method for constructing a thesaurus of real-world entities using the semantic similarity computed from the web directory. For each entity, the thesaurus lists the entities most similar to the target entity in different aspects. The thesaurus can be used in many applications such as entity disambiguation and query expansion [11]. Second, we built a categorization of entities by clustering them based on their semantic similarities. The categorization is at much finer level than common named entity recognition, which is limited to a small set of broad categories. The fine grained categorization of entities is useful for various applications such as semantic annotation of web pages [2]

There is some related work concerning measuring semantic similarity of entities. In social network extraction, co-occurrence analysis is widely used to detect groups of people. For example, Flink [8] and Polyphonet [7] both utilize the hit counts returned by search engines for individual names and their conjunctions (i.e. "X AND Y" and "X or Y"). Bollegala et al. [1] addresses the problem of measuring semantic similarity using a combination of hit counts and text snippets. Compared to [1, 7, 8], our method utilizes a web directory rather than general web search engines. The categories (such as "Science > Physics > History > People > Einstein, Albert") provide explicit information about the properties of the entigy. Since they are specified in a controlled vocabulary, they are much less noisy than general web pages.

The remainder of the paper is organized as follows. In Section 2, we present the proposed method. We then describe the method for constructing a thesaurus of entities and show the thesaurus we built for the Forbes 2000 in Section 3. Section 4 presents the method and experiment for fine grained categorization of entities.

2. Proposed method

Google Directory provides an effective interface to the ODP web directory. When queried with the name of an entity, Google Directory will return the web pages most relevant to the entity along with their categories in the directory hierarchy. Figure 1 illustrates the top results for the entity "IBM T.J. Watson Research Center" returned by Google Directory. In the search result, the *directory heading* that denotes the category of the web page (such as "Science > Math > Research > Institutes") gives critical information about the features of the entity

Figure 1. Exemplar search results

"Science Math Research Institutes Computers Internet Resources Research Science Technology Nanotechnology Research Institutions Computers Companies IBM"

Figure 2. Exemplar profile

Thus, semantic similarity of entities can be measured by comparing their corresponding directory headings in the web directory.

A straightforward method to measure similarity in a taxonomy is to count the edges of the shortest path connecting the two entities in the taxonomy. This method assumes that all links in the taxonomy represent uniform distance. However, close inspection of the implicit taxonomy of the web directory reveals that the hierarchy is sometimes mistaken. As noted by Uschold [12], the relation between the categories is a mix of various relationships such as specialization and part-of. Thus the traditional edge-counting method is not suitable for measuring similarity in web directory.

Instead, we can interpret the categories of the web pages relevant to an entity from a different perspective. The terms in the directory headings actually provide explicit information about the important features of the corresponding entity, such as location and industry of a company. When composed together, the directory headings produce a *profile* of the entity in a controlled vocabulary. The semantic similarity between the entities can be measured by comparing their profiles built from the directory headings. To implement this idea, we extract the directory headings of the top pages relevant to entity e, remove the category-subcategory boundary symbol (i.e. ">"), and concatenate them into a text snippet. The text snippet is the profile $P(e)$ of the entity e. The profile for "IBM T.J. Watson Research Center" is illustrated in Figure 2.

To compare the profile of entities, we adopt the traditional vector space model in information retrieval. Each text snippet p is converted into a vector v_p of the terms contained in p. In weighing the terms in p, three factors are taken into account:

- rank of the page containing the term in its directory heading, in the search results returned by the Google Directory. The rank indicates how relevant the web page is to the entity;
- level of the category containing the term in the directory hierarchy. Terms at higher levels are mostly generic and basic features of the entity.
- frequency of occurrences, total number of times the stemmed term occurs in the profile, which indicates the salience of the feature of the entity;

The three factors are implemented in a single metric of term weighting. It works as follows. In generating the profile from directory headings of the relevant web pages, the order of the pages in the search result and the order of the categories in the hierarchy are preserved. Consequently, terms appearing near the beginning of the profile are terms in more relevant pages and in higher-level categories. Accordingly, the occurrence of the first term in the profile is assigned a static maximum score. The scores of the other term occurrences in the profile decrease linearly. The last term occurrence in the profile is assigned a static minimum score. Finally, the weight of a term is the sum of the scores of all its occurrences in the profile.

Using the vector representation v_p of profile $P(e)$, the semantic similarity of entities are computed as the cosine similarity of the corresponding vectors. It should be noted that the similarity measurement based on term vectors is not much affected by synonyms because the terms are driven from a controlled vocabulary.

In addition, the terms shared by the two vectors v_{p1} and v_{p2} indicate the shared features of the two entities. Those terms are returned as the *commonality keywords* to represent the specific aspects of similarity between the two entities. Consequently, the measurement of semantic similarity of entities returns two values: similarity score and commonality keywords.

3. Thesaurus of named entities

Measuring the semantic similarity of named entities plays an important role in many applications such as query expansion [11] and entity disambiguation. For common words, general thesauri such as WordNet have proven to be helpful for query expansion and word sense disambiguation. However, these general thesauri provide very limited coverage of entities, especially for emerging entities. A thesaurus that lists
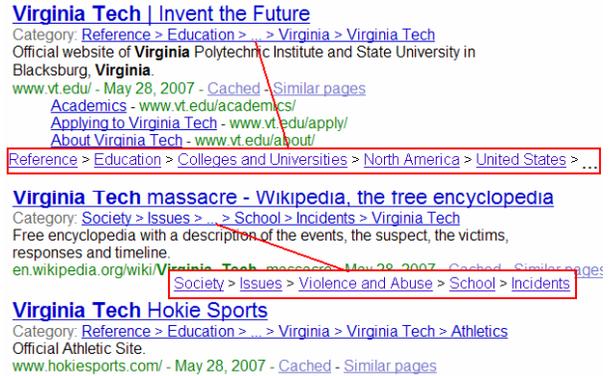

Figure 3. Top results for "Virginia Tech"

the similar entities of real-world entities will be beneficial for these applications.

In general thesauri, semantic similarities of words are measured and grouped in different senses. Similarly, similarities of named entities lie in different dimensions. For example, "Virginia Tech" and "Yale University" are similar in the sense that they are both American universities. On the other hand, "Virginia Tech" and "Columbine High School" are similar in the sense that they are the sites of school massacres. A thesaurus of entities should capture the similarities between entities in different dimensions.

Because our method for measuring semantic similarity is based on the web directory, it can benefit from the huge amount of web content. On the web, there are various web pages characterizing different aspects of an entity. When indexed by the web directory, the pages about the distinct properties of the entities are assigned into different categories. Figure 3 shows the top results for "Virginia Tech" in Google Directory. When compared with "Yale University", the terms in the directory heading of the first result will be matched (i.e. "education", "college", "university", etc.); when compared with "Columbine High School", the terms in the directory heading of the second result will be matched (i.e. "violence", "abuse", "incidents", etc). Using the directory headings of different web pages about the entities, the semantic similarities between named entities can be measured and highlighted in different aspects.

As described in Section 2, the measurement of semantic similarity of entities returns a similarity score as well as a set of commonality keywords representing the shared attributes of the entities. With the similarity scores and commonality keywords, we can automatically construct a thesaurus of entities in the style of WordNet. Given a list of entities, we first compute the semantic similarity between each pair of entities. For each entity e, the similar entities e_i of e with the same commonality keyword c can be grouped

"Isuzu Motors":	*Autos*	("Hyundai Motor", 0.82), ("General Motors", 0.75), ("Mitsubishi Motors", 0.75) ("Porsche", 0.75) ("Honda Motor", 0.74)…
	Japan	("Idemitsu Kosan", 0.69), ("Toyota Industries", 0.67), ("Suzuki Motor", 0.54), ("Hanwa", 0.45), ("Daimaru", 0.44)…
"Sony":	*Entertainment*	("News Corp", 0.58), ("Liberty Media Holding", 0.54), ("Viacom", 0.53), ("Vivendi", 0.53), ("Fujifilm Holdings", 0.38), ("Walt Disney", 0.38)…
	Hardware	("Toshiba", 0.46), ("Mitsubishi", 0.42), ("Linear Technology", 0.40), ("SanDisk", 0.39)…
	Electronics	("Best Buy", 0.50), ("Garmin", 0.47), ("Toshiba", 0.46), ("Samsung", 0.43)…
"Microsoft":	*Software*	("Fanuc", 0.62) ("Adobe Systems", 0.59) ("Apache", 0.58) ("Oracle", 0.58) ("SAP", 0.57) ("Oki Electric Industry", 0.56), ("Wipro", 0.52) ("Infosys Technologies", 0.50) …
	Databases	("Oracle", 0.58), ("Pitney Bowes", 0.32)…
	WWW	("Apache", 0.584112)…

Figure 4. Thesaurus of real-world entities

together and ordered with their similarity scores s_i. The entries in the thesaurus are in the following format,

e: c_1 - (e_{11}, s_{11}), (e_{12}, s_{12}), (e_{13}, s_{13}) …
c_2 - (e_{21}, s_{21}), (e_{22}, s_{22}), (e_{23}, s_{23}) …
…

In our experiment, we used the list of companies from the Forbes 2000 [4], a list of the world's 2000 largest public companies. Figure 4 shows the thesaurus entries for "Isuzu Motors", "Sony" and "Microsoft". The thesaurus lists the most similar entities to the target entity in different aspects. For "Isuzu Motors", which is a Japanese automobile company, the thesaurus lists other automakers under the commonality keyword *Autos* and other Japanese companies under the commonality keyword *Japan*. For "Sony" and "Microsoft", the thesaurus lists the different businesses involving the company and the similar companies in those businesses. Close observation reveals that the concepts denoted by the commonality keywords are not exclusive to each other. For instance, *Database* is a subclass of *Software*. In future work the semantics of web directory taxonomy will be further explored to create better presentation for the thesaurus.

4. Fine grained categorization of entities

Named Entity Recognition (NER) is vital in many language and information processing applications. Current NER tools classify entities into broad categories such as person, organization, and location. However, for complex applications concerning semantics, such as semantic annotation of web pages [2], fine grained categorization of entities is necessary.

There is some research on the subcategorization of named entities in the area of computational linguistics and information retrieval. Fleischman and Hovy [3] propose a supervised learning method for classifying named entities denoting people. Their classifiers use various features including local context and topic signatures. Pasca [10] presents a method for acquiring named entities in arbitrary categories from web

documents. The method applies lexico-syntactic extraction patterns which is initially hand-built and learned incrementally in the processing of the corpus.

Our method for subcategorizing named entities takes a different approach from [3, 10]. It does not require any text corpus. Instead, a list of named entities in a coarse category is clustered into subcategories according to their semantic similarity computed with the web directory. The list of named entities in a coarse category can be obtained by combining existing gazetteer lists or generated with traditional NER tools. Based on semantic similarities between the entities, the entities can be clustered into semantically coherent clusters. Moreover, the commonality keyword with the highest total weight in the cluster can be used as the label of the cluster, which denotes the most salient feature of the entities in the cluster.

In order to evaluate the performance of the proposed method in categorizing named entities, we collected named entities from Wikipedia [13] in three coarse categories: *company*, *people*, and *city*. The list of companies was created from 8 lists of companies in different industry; the list of people consisted of Nobel Prize winners in 5 fields; and the list of cities was compiled from 6 lists of cities in different countries. Table 1 presents the details of the three lists of entities.

Each entity was queried in Google Directory and the directory headings of the top 3 pages were extracted. The entities without any search results were deleted from the lists. As shown in Table 1, the web directory provided a large coverage of the entities. For each coarse category, we computed the semantic similarity between every pair of entities in the category. We used group-average agglomerative hierarchical clustering [6] to cluster the entities in each coarse category into the respective number of subcategories, i.e. 8 clusters for *company*, 5 clusters for *people*, and 6 clusters for *city*.

We employed Rand Index [6] to evaluate the clustering results. The subcategorization by the original Wikipedia lists was used as ground truth. For every pair of the entities in the coarse directory, if they

Table 1. Lists of entities in coarse categories

Coarse category	Subcategories	Total No. of entities	Percentage of entities without search results
company	Advertising, Automobile, Computer and video game, Fast-food, Airlines, Department stores, Petroleum, Software	834	2.9%
people	Economics, Physiology or Medicine, Chemistry, Physics, Literature	636	12.9%
city	Australia, Canada, China, Japan, Russia, South Africa, USA	734	7.9%

were in different subcategories in Wikipedia but were assigned to the same cluster, the pair was considered as *false positive*; if they were in the same subcategory in Wikipedia but were assigned to different clusters, the pair was considered as *false negative*. Rand Index penalizes both false positive and false negative.

$$RI = \frac{N_{True\ Possitive} + N_{True\ Negative}}{N_{True\ Possitive} + N_{True\ Negative} + N_{False\ Possitive} + N_{False\ Negative}}$$

Table 2. Performance of entity subcatigorization

Coarse category	Rand Index
company	84.0%
people	77.8%
city	87.5%

Table 2 reports the Rand Index for the three lists of entities. Among the three coarse categories, *city* achieved the highest Rand Index. This is because web pages about locations are usually unambiguously categorized with their geographic feature. In comparison, web pages about a company or a person are assigned into different categories according to the different features of the entity. Clustering of *people* turned out to be noisier than *company* and *city* because of the problem that people have same names. The problem can be address if we know more information about the people to help disambiguation.

5. Conclusion

In this paper, we propose a method for measuring semantic similarity between named entities. The method exploits the semantic knowledge embedded in the Open Directory Project, the most comprehensive human edited hierarchy of web documents. By querying Google Directory with the name of an entity, we can locate the web pages relevant to the entity, which capture various features of the entity. Using the directory headings assigned to the relevant pages, we can automatically create a profile of the entity. The semantic similarity between entities is measured based on their profiles. We apply the measurement of semantic similarity to two knowledge acquisition tasks. First, we present a method for automatically constructing a thesaurus of entities. The thesaurus lists the entities similar to the target entity in different aspects. Second, we built subcategorizations of entities by clustering them based on their semantic similarities. Results of our experiments show that the proposed method can effectively capture the semantic similarity between real-world entities. In future research, we will further explore the semantics of the web directory taxonomy in measuring semantic similarity. Our method may also be combined with other methods based on co-occurrence and local context analysis to better capture the relationships between entities.

6. References

[1] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring Semantic Similarity between Words Using Web Search Engines", *Proc. 16th World Wide Web Conference*, 2007.

[2] P. Cimiano, S. Handschuh, and S. Staab. "Towards the self-annotating web", *Proc. 13th World Wide Web Conference*, 2004.

[3] M. Fleischman and E. Hovy, "Fine grained classification of named entities", *Proc. Conference on Computational Linguistics*, 2002.

[4] Forbes 2000: http://www.forbes.com/

[5] Google Directory: http://directory.google.com/

[6] C.D. Manning, P. Raghavan, and H. SchÄutze. *An Introduction to Information Retrieval*, Cambridge University Press, 2005.

[7] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: An advanced social network extraction system", *Proc. 15th World Wide Web Conference*, 2006.

[8] P. Mika, "Flink:semantic web technology for the extraction and analysis of social networks", *Journal of Web Semantics* 3(2), 2005

[9] Open Directory Project: http://dmoz.org/

[10] M. Pasca, "Acquisition of categorized named entities for web search", *Proc. 13th ACM conference on Information and knowledge management*, 2004

[11] M. Sahami and T. Heilman. A web-based kernel function for measuring the similarity of short text snippets. *Proc. 15th World Wide Web Conference*, 2006.

[12] M. Uschold and R. Jasper, "A Framework for Understanding and Classifying Ontology Applications", *Proc. of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends*, 1999

[13] Wikipedia: http://wikipedia.org/