Extracting Semantic Knowledge from Wikipedia Category Names

First Author
Institute 1
Lane 1
Country 1
email1@organization.com

Second Author
Institute 2
Lane 2
Country 2
email2@organization.com

ABSTRACT

Wikipedia being a large, freely available, frequently updated and community maintained knowledge base, has been central to much recent research. However, quite often we find that the information extracted from it has extraneous content. This paper proposes a method to extract useful information from Wikipedia, using Semantic Features derived from Wikipedia categories. The proposed method provides improved performance over the state of the art Wikipedia category based method. Experimental results on benchmark datasets show that the proposed method achieves a correlation coefficient of 0.66 with human judgments. The Semantic Features derived by this method gave good correlation with human rankings in a web search query completion application.

1. INTRODUCTION

Wikipedia[1] is a free online encyclopedia that is constructed in a collaborative effort of voluntary contributors. With the content being frequently updated and coverage growing exponentially, Wikipedia as a knowledge base has been central to much recent research in Information Extraction[7] and Knowledge base construction[23, 26]. However extracting semantically rich knowledge from it continues to be a difficult task. Several approaches[22, 28, 23, 11] to extract information from Wikipedia, use supervised machine learning using Wikipedia's Category hierarchy. This paper presents a method with six Semantic Features for this information extraction.

Wikipedia articles form a network of semantically related terms, while the categories are organized in a taxonomy-like structure called Wikipedia Category Network (WCN)[28]. In categorizing an article under a category, human judgment is involved. It is this human judgment that we are trying to capitalize upon analyzing category names with respect to the articles they categorize. We analyze the entire collection of Wikipedia category names along with the titles of the articles they categorize. We analyze it for semantic and hierarchical information, and derive the Semantic Features proposed in this paper.

We evaluate the performance of these Semantic Features on standard datasets. We use the Semantic Features in a machine learning based semantic relatedness prediction system for evaluation. Since the datasets are limited in size, we also evaluate the Semantic Features by applying them to a web query completion task.

The contributions of this paper are

1. We propose **six Semantic Features** that can be used in supervised machine learning algorithms which extract information from semi-structured knowledge bases like Wikipedia.
2. We propose methods to automatically derive **semantically related word pairs** from Wikipedia Category names.

2. RELATED WORK

Semantic relatedness indicates how much two concepts are related in a taxonomy by using all relations between them (including hyponymic/hypernymic, meronymic and functional relations). Semantic relatedness measures are used in many applications like word sense disambiguation [16] and information retrieval[4].

Past decade has seen several new techniques which infer semantic relatedness from the structure and content of Wikipedia. With over four million articles and still growing, Wikipedia is one of the largest encyclopedias. With the articles being categorized into a network of categories, it also contains a wealth of explicitly defined semantics. This rare combination of scale and structure has been used by studies on Wikipedia Category Network(WCN). Some of the prominent ones include WikiRelate[22], where semantic relatedness measures proved on WordNet [2] were modified and applied to suit Wikipedia. WikiRelate is considered as state of the art in Wikipedia category based semantic relatedness measures[13]. We use the measurements from WikiRelate as the baseline for our current study. Zesch & Gurevych[29] measures the semantic relatedness in terms of the path length between the categories in the hierarchy.

Every Wikipedia article contains a list of categories it belongs to. The ordering of categories in this category list, conveys the overall significance of the category to the article. Further the categories to the left in the list are of higher

[1]http://www.wikipedia.org

[2]http://wordnet.princeton.edu/

AKBC 2013, San Francisco, USA

significance [5]. So we measure the average leftness of a category. This is the basis of the 'Leftness' Semantic Feature proposed in this paper. We also measure the ordering of the category in the category list and relative difference in ordering between two categories. This is the basis of the 'Similarity' and 'Relative Popularity' Semantic Features proposed in this paper.

3. APPROACH

Our approach is **to analyze the WCN** for hierarchical and semantic information, **and derive features from it**. We start by analyzing the taxonomic aspects of WCN. The hierarchical aspects of a category like number of children(articles), number of sibling categories and the linkage between them is translated into Semantic Features. Then the syntactic aspects of the category are analyzed to see how the category names are broken down to form word pairs.

WordNet is a general-purpose lexical ontology that groups semantically related words (i.e. synsets) for concepts[2]. It is created by linguists and has a strictly defined hierarchy for words[29]. In comparison, WCN is not created by linguists and the hierarchy is criticized to be flat[26] and containing cycles[3]. However Wikipedia far out performs WordNet in terms of word coverage. To combine the advantages of both, we use WCN as the basis of our study while avoiding usage of its full category hierarchy.
Earlier works on WCN [22][29], consider all paths between the categories in the hierarchy. Our method only considers path between two categories through an article. This is done by analyzing selective parts of WCN as explained in Figure 1. In the figure, the categories 'Domesticated Animals' and 'Pet Mammals' are connected through three common articles('Cat','Dog' and 'Ferret'). The fact that there are three common articles between the two categories indicates relatedness between the categories. Similarly the category 'Domesticated Animals' has two articles('Cat' and 'Dog') under it. This is indicative of relatedness between words "*domestic*" and "*animal*". Analyzing the WCN this way helps in extracting useful information with less-noise.

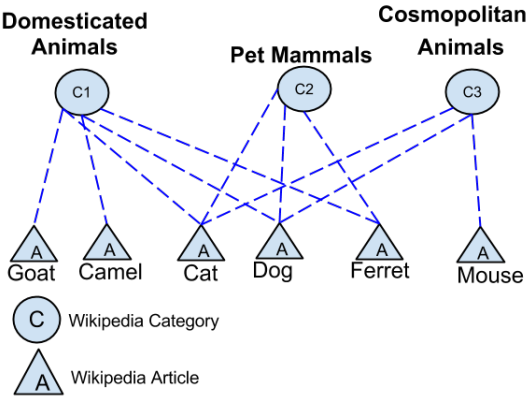

Figure 1: Wikipedia Category Network(WCN)

In this paper we derive semantically related word pairs from Wikipedia Category names. The category names themselves hold a wealth of semantic information as they are manually generated and are used to gather concepts with similar properties. In a previous work Schonhofen[21] finds Wikipedia category names more useful to describe documents than the respective full text in general scenarios.
Our dataset is a collection of Wikipedia category names with the articles' titles they categorize[3]. We analyze the WCN for semantic and hierarchical information. Preprocessing was done to remove Wikipedia administrative categories and some of the very large size categories like 'Living people' from the dataset. We analyze this dataset for hierarchical and syntactic aspects.

3.1 Heirarchical analysis

Here we analyze a category with respect to its position in the WCN, first for a single category and then for a category pair.

3.1.1 Single category

NormalizedRepresentation($\mathbf{NR_1}$) : The category name represents the articles for a concept or topic. So the number of articles under a category is measured as its representation[4]. To avoid bias to large categories we normalize this measure by dividing it by the sum of representations of the categories under which articles of the given category fall. With reference to the WCN in Figure 1 this can be written as

$$NR_1(C1) = \frac{\#article(C1)}{\#article(C1) + \#article(C2) + \#article(C3)} \quad (1)$$

where *#article (C1)* is number of articles in category *C1*, *#article (C2)* is number of articles in category *C2* and so on.
Similarity : Gyllstrom & Moens[5] defines categorical similarity as the mean similarity of a category's articles among each other. We calculate Similarity of a category as the sum of similarity between articles. Similarity between articles A1 and A2 is given by

$$similarity(A1, A2) = \frac{|Categories_{A1} \cap Categories_{A2}|}{|Categories_{A1} \cup Categories_{A2}|} \quad (2)$$

where $Categories_{Ai}$ represents category list of article Ai.
Leftness : Leftness of a category (represented as **$\mathbf{Leftness_1}$**), indicates relevance of the category to the article[5]. Let there be N articles in which category C figures. C's leftness is given by

$$Leftness_1(C) = \sum_{i=1}^{N} \frac{\#categories(Ai) - pos(C)}{\#categories(Ai)} \quad (3)$$

where *#categories(Ai)* is number of categories in article Ai and *pos(C)* is the position of C in the Ai's category list.

3.1.2 Category Pair

For every category pair we count the common articles(M). We measure the position of the individual categories in the category lists of common articles. Normalizing these we get the *Normalized Representation* and *Relative popularity (RP)*, as explained below.
NormalizedRepresentation($\mathbf{NR_2}$) :

$$NR_2(C1, C2) = \frac{\#article(C1, C2)}{\#article(C1) + \#article(C2)} \quad (4)$$

[3]We use only category names and article titles collected from Wikipedia as of June 2013. Total of 4.26 million titles.
[4]The name *representation* was chosen since the value showed the number of articles representing the category or pair.

where *#article (C1,C2)* is number of articles between categories *C1* and *C2*.

Relative Popularity(RP) : The number of positions that separates two categories in the article's category list is normalized with the total number of categories present in the article's category list. This is called RP. The aggregate of RP over all common articles of the category pair is called as RP of the category pair.

$$RP(C1, C2) = \sum_{i=1}^{M} RP_i \quad (5)$$

Leftness$_2$: Leftness of a Category pair is calculated as the minimum of $Leftness_1$ of the member categories. $Leftness_1$ is derived with Equation 3.

3.2 Syntactic analysis

Category names are phrases. We analyze the Part-of-Speech (POS) tags of phrases to identify any POS tag patterns. Category names are found to follow either of the two patterns, *NounPhrase* or *NounPhrase*(IN *NounPhrase*)+. The *NounPhrase* pattern includes any of the POS tag in the set { NN, NNP, NNPS, NNS, VBN, VBZ, VBD, VB, CD, DT, JJ , CC} with tag repetitions. IN stands for any of the set {by, in, from, for, out, of, with, at, about} and + indicates repetitions.

Category names are broken down to create word pairs. Category names of the pattern *NounPhrase* is split from multiword to single word, like: '`Guinea pig`' is broken into '`guinea`' and '`pig`'. The single words are then paired as in '`guinea_pig`'. Category names of the pattern *NounPhrase*(IN *NounPhrase*)+ has semantic knowledge bound in itself. They are broken at preposition(i.e IN) leading to words from first *NounPhrase* getting paired with words from second *NounPhrase*. For example '`Cats in art`' and '`Films about cats`' into '`cats_films`', '`cats_art`' and '`art_films`'.

4. EXPERIMENTS

The Semantic Features are evaluated by the improvements they bring to Semantic Relatedness measurements. We consider four standard datasets for semantic relatedness evaluation namely Miller & Charles [12], Rubenstein & Goodenough [20], WordSimilarity-353 Test Collection [4] and Temporal Semantic Analysis [18]. These are hereafter referred to as **MC30**, **RG65**, **WS353** and **TS287** respectively. Following the literature on semantic relatedness, we evaluate the performance by taking the Pearson product-moment correlation coefficient(r) between the relatedness measure scores and the corresponding human judgments. For WS353 and TS287 datasets we report the correlation, computed first with standard (WordNet based) relatedness measures alone and then with Semantic features added. The WS353 and TS287 datasets are large enough to be partitioned into training and testing[22]. For each word pair in these datasets, the WordNet similarity measures specified in Table 1 are computed using the implementation available in WORDNET::SIMILARITY [17] package. Along with the WordNet similarity measures Table 1 also shows the correlation of each measure on WS353 and TSA287 datasets. We find that the best correlation is given by Extended gloss overlap[1] measure[5].

The WordNet measures are integrated by performing regression using a Support Vector Machine [24] . This helps estimate the functional dependence of the human relatedness judgments on multiple relatedness scores. The learner was trained and tested using all WordNet measures in Table 1. The experiments was repeated adding Semantic Features to the WordNet measures. The learner was trained and tested using WordNet scores and Semantic features. We used an RBF kernel with a 5 fold cross-validation, on the training data.

The MC30 and RG65 datasets were small compared with WS353 and TS287 datasets. So we evaluated it as follows. For each word pair in these datasets, the best performing WordNet measure (refer Table 1) was combined with the Semantic Feature values to arrive at a semantic relatedness measurement. The correlation of this measurement on MC30 and RG65 datasets is reported in Table 2.

[5]In all tables, best results are highlighted in bold

Table 1: WordNet similarity measures correlation

Measure	r(WS353)	r(TSA287)
Jiang & Conrath(1998)[8]	0.02	0.08
Resnik (1995)[19]	0.01	0.08
Lin [10]	0.01	0.02
Leacock & Chodorow[9]	0.02	0.02
Wu & Palmer[27]	0.01	0.02
Hirst & St-Onge[6]	0.35	0.29
Patwardhan et al.[17]	0.39	0.32
Banerjee & Pederson[1]	**0.43**	**0.42**
Patwardhan [15]	0.33	0.27

5. RESULTS

Table 2: Correlation with human judgments on SR Evaluation datasets

Word Pair Source	Feature	MC30	RG65	WS353	TSA287
WordNet(WN)		0.33	0.33	0.37	0.40
Single Category	$WN + NR_1$	0.31	0.4	0.57	0.51
	WN+Similarity			0.57	0.48
	WN+Leftness			0.55	0.49
	Combined			0.57	0.48
Category Pair	$WN + NR_2$	0.29	0.49	0.44	**0.66**
	WN+RP			0.21	0.37
	WN+Leftness			0.31	0.21
	Combined			0.34	0.61

Table 2 shows the correlation coefficients of the different Semantic Features with human judgments. The experimental setup is similar to that of Strube & Ponzetto[22] barring the differences in WCN usage and not using article text or disambiguation pages. The correlation of 0.37 on WordNet measures on WS353 dataset is also in line with results of [22]. The Semantic Features NR_1, RP, $Leftness_1$, NR_2, Similarity and $Leftness_2$ were tested in combination with WordNet features. The best performance was **0.66**(statistically significant, two tailed t-test with $\alpha = 0.05$). This was given by NR_2 in combination with WordNet features. This suggests that category's representation of articles can be used as an indicator of semantic relatedness between words. While most of word pairs in TSA287 and WS353 was present in the knowledge base, coverage of RG65 and MC30 was less. This could be because TSA287 and WS353 are temporally near to Wikipedia corpus we used, compared to RG65 and MC30. This temporal aspect of word coverage is explained in [18].

Table 3: Rank correlation in search query completion task.

Semantic Feature		Correlation(τ)
Single Category	NR_1	0.27
	Similarity	**0.56**
	$Leftness_1$	0.27
Category Pair	NR_2	0.31
	RP	**0.42**
	$Leftness_2$	0.12
Human		1.00

6. EVALUATION

We evaluate the Semantic Features by judging its performance in a text processing application as suggested in [14, 29]. We use it in the search query completion problem. The problem consists of a target word and four candidate words or phrases. The objective is to rank the candidates in the decreasing order of their probability to complete the target word in a web search query. In other words, the candidate words are ordered in the decreasing order of relatedness to the target word and the candidate word with the maximum relatedness value is ranked first.

We used the revised version of General Service List of English Words (GSL)[25], provided by John Bauman[6] with 2284 words. These words act as target words. Candidate words are generated using Semantic Feature values. Due to the similarity of the problem to the 'Word Choice Problem' described in [29], the test setup specified there was followed. Barring few words, almost all words in the GSL was found in the knowledge source, and candidate list was generated for them. In order to avoid bias, we randomly chose a subset of 100 words from the GSL as our problem set. For these words, list of all possible candidates in the knowledge source is created. The problem set and candidate list is then evaluated by human evaluators as explained below.

Ten human judges[7] were presented with the problem set. The judges were researchers in the 21 to 30 years age group with near-native speaker proficiency in English language. Scores from human judges were merged into one. This was possible owing to high inter annotator agreement and similar background knowledge (all are researchers in the same lab) of the human judges. Outlier target words, where consensus could not be reached, was excluded from problem set. Judges were asked to pick the top four candidates. On this list the rankings provided by all the six Semantic Features proposed in this paper was evaluated. Correlation between the rankings produced by each Semantic Feature with that of (combined) human rankings is measured using Kendall's tau (τ) coefficient given in Table 3. Best correlation was obtained for RP and Similarity features.

7. DISCUSSION

We analyze the strengths and weaknesses of our method in the contexts of two applications namely Search query completion and Word choice problem. Search query completion is explained in previous section. Here the system's ability to rank the candidate words based on their probability to complete the target word in a web search query is evaluated. In the search query completion task, we see better performance with co-occurring words like '`domesticated_mammal`' in Figure 1. In the extracted words frequency of co-occurring words are high. This can be attribute to folksonomy, as Wikipedia is created by a community. While the words '`domesticated`' and '`mammal`' are related, the relation is not meaning or synonym. Such related word pairs suit Search query completion application.

The second application we studied is the Word choise problem. It consists of a target word and four candidate words or phrases. The objective is to pick the candidate word that is most closely related to the target. The system performs this by computing the semantic relatedness measure between the target and each of the candidate words. The candidate with the maximum semantic relatedness value is chosen.

We found that the proposed Semantic Features gives better results in Search query completion application than Word choice application. There could be many reasons for this. First of all, the words extracted by our method is predominantly nouns, while most of the target words in Word choice application were verb. Secondly, it is unlikely that a Wikipedia category name has two words of similar meanings. Hence it was difficult to find words paired with their meaning or synonyms in the extracted word corpus. To solve Word choice problems, meanings and synonyms was needed. Hence the proposed Semantic Features does not suit Word choice problem.

Thus we see that the proposed Semantic Features can be used in applications that need to predict semantic relatedness between words, especially between frequently co-occurring nouns.

8. CONCLUSIONS

In this paper, we propose a method to derive Semantic Features. These features can be used in supervised machine learning algorithms that calculate semantic relatedness between two given words or named entities. The method consists of six semantic features that can be automatically extracted from the Wikipedia Category Network. We integrated these features with established WordNet based semantic relatedness measures using support vector machines. Training and testing was done using four semantic relatedness evaluation data sets. Proposed method outperformed the baselines established using Wikipedia Category Network on a benchmark dataset. It achieved good correlation (correlation co-efficient of 0.66) with human ratings for semantic relatedness on the benchmark dataset. It only processes the Wikipedia Category names and the title names associated with them (Downloading of article pages is not necessary).

A contrasting feature of our method compared to the other methods using Wikipedia Category Network is that our method does not require the entire Wikipedia Category hierarchy. Instead it uses only selective parts of it, doing a shallow analysis of the hierarchy. Therefore, the proposed method can be applied in many tasks where such deep and well defined taxonomies do not exist or are not up-to-date. We employed the proposed method in a Web query completion experiment. Results of our experiments indicate that the proposed method can capture semantic relatedness between nouns. In future research, we intend to apply the proposed Semantic Features in word clustering and query suggestion applications.

[6] http://jbauman.com/aboutgsl.html

[7] Eight men and two women

9. REFERENCES

[1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, 2003.

[2] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th international conference on World Wide Web*, pages 757–766, New York, NY, USA, 2007. ACM.

[3] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantic relationships between wikipedia categories. In *In 1st International Workshop: âĂIJSemWiki2006 - From Wiki to SemanticsâĂİ (SemWiki 2006), co-located with the ESWC2006 in Budva*, 2006.

[4] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: the concept revisited. In *WWW*, pages 406–414, 2001.

[5] K. Gyllstrom and M.-F. Moens. Examining the "leftness" property of wikipedia categories. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2309–2312, New York, USA, 2011. ACM.

[6] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms, 1997.

[7] E. H. Hovy, R. Navigli, and S. P. Ponzetto. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27, 2013.

[8] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, pages –1–1, 1997.

[9] C. Leacock, G. A. Miller, and M. Chodorow. Using corpus statistics and wordnet relations for sense identification. *Comput. Linguist.*, 24(1):147–165, Mar. 1998.

[10] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[11] Q. Liu, K. Xu, L. Zhang, H. Wang, Y. Yu, and Y. Pan. Catriple: Extracting triples from wikipedia categories. In *Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web*, ASWC '08, pages 330–344, Berlin, Heidelberg, 2008. Springer-Verlag.

[12] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. volume 6, pages 1–28. Psychology Press, 1991.

[13] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *In Proceedings of AAAI 2008*, 2008.

[14] A. Panchenko and O. Morozova. A study of hybrid similarity measures for semantic relation extraction. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, HYBRID '12, pages 10–18, Stroudsburg, PA, USA, 2012. ACL.

[15] S. Patwardhan. Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. Master's thesis, University of Minnesota, Duluth, August 2003.

[16] S. Patwardhan, S. Banerjee, and T. Pedersen. Senserelate::targetword: a generalized framework for word sense disambiguation. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, ACLdemo '05, pages 73–76, Stroudsburg, PA, USA, 2005. ACL.

[17] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA, 2004. ACL.

[18] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 337–346, New York, NY, USA, 2011. ACM.

[19] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[20] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, Oct. 1965.

[21] P. Schonhofen. Identifying document topics using the wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 456–462, Washington, DC, USA, 2006. IEEE Computer Society.

[22] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, pages 1419–1424. AAAI Press, 2006.

[23] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[24] V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[25] M. West. In *A General Service List of English Words.* London: Longman, Green and Co, 1953.

[26] F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 41–50, New York, NY, USA, 2007. ACM.

[27] Z. Wu and M. Palmer. Verb semantics and lexical selection. 1994.

[28] T. Zesch and I. Gurevych. Analysis of the wikipedia category graph for nlp applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, pages 1–8, Rochester, Apr. 2007. Association for Computational Linguistics.

[29] T. Zesch and I. Gurevych. Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words. *Nat. Lang. Eng.*, 16(1):25–59, Jan. 2010.