A Survey of Exploiting WordNet in Ontology Matching

Feiyu Lin and Kurt Sandkuhl

Abstract Nowadays, many ontologies are used in industry, public adminstration and academia. Although these ontologies are developed for various purposes and domains, they often contain overlapping information. To build a collaborative semantic web, which allows data to be shared and reused across applications, enterprises, and community boundaries, it is necessary to find ways to compare, match and integrate various ontologies. Different strategies (e.g., string similarity, synonyms, structure similarity and based on instances) for determining similarity between entities are used in current ontology matching systems. Synonyms can help to solve the problem of using different terms in the ontologies for the same concept. The WordNet thesauri can support improving similarity measures. This paper provides an overview of how to apply WordNet in the ontology matching research area.

1 Introduction

The Semantic Web provides shared understanding, well structured content and reasoning for extending the current web. Ontologies are essential elements of the semantic web. Nowadays, many ontologies are used in industry, public adminstration and academia. Although these ontologies are developed for various purposes and domains, they often contain overlapping information. To build a collaborative semantic web, which allows data to be shared and reused across applications, enterprises, and community boundaries [22], it is necessary to find ways to compare, match and integrate various ontologies.

Ontology matching in general is based on finding similar entities in the source ontologies or finding translation rules between ontologies. Different strategies (e.g.,

Feiyu Lin
Jönköping University, Jönköping, Sweden, e-mail: `feiyu.lin@jth.hj.se`

Kurt Sandkuhl
Jönköping University, Jönköping, Sweden, e-mail: `kurt.sandkuhl@jth.hj.se`

Please use the following format when citing this chapter:

Lin, F. and Sandkuhl, K., 2008, in IFIP International Federation for Information Processing, Volume 276; *Artificial Intelligence and Practice II*; Max Bramer; (Boston: Springer), pp. 341–350.

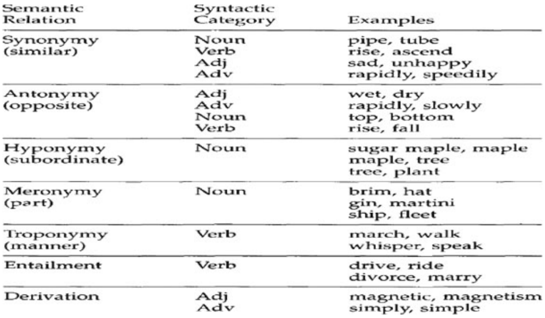
string similarity, synonyms, structure similarity and based on instances) for determining similarity between entities are used in current ontology matching systems. When comparing ontology entities based on their labels, synonyms can help to solve the problem of using different terms in the ontologies for the same concept. For example, an ontology might use "diagram", another ontology could use "graph" referring to the same concern.

The WordNet[25] can support improving similarity measures. This paper provides an overview of how to apply WordNet in the ontology matching research area.

2 WordNet

WordNet is based on psycholinguistic theories to define word meaning and models not only word meaning associations but also meaning-meaning associations [7]. WordNet tries to focus on the word meanings instead of word forms, though inflection morphology is also considered. WordNet consists of three databases, one for nouns, one for verbs and a third for adjectives and adverbs. WordNet consists of a set of synonyms "synsets". A synset denotes a concept or a sense of a group of terms. Synsets provide different semantic relationships such as synonymy (similar) and antonymy (opposite), hypernymy (superconcept)/hyponymy (subconcept)(also called Is-A hierarchy / taxonomy), meronymy (part-of) and holonymy (has-a). The semantic relations among the synsets differ depending on the grammatical category, as can be seen in Figure 1 [11]. WordNet also provides textual descriptions of the concepts (*gloss*) containing definitions and examples. WordNet can be treated as a partially ordered synonym resources.

Fig. 1 Semantic relations in WordNet. (Source: [11])



Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	Noun Verb Adj Adv	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Adj Adv Noun Verb	wet, dry rapidly, slowly top, bottom rise, fall
Hyponymy (subordinate)	Noun	sugar maple, maple maple, tree tree, plant
Meronymy (part)	Noun	brim, hat gin, martini ship, fleet
Troponymy (manner)	Verb	march, walk whisper, speak
Entailment	Verb	drive, ride divorce, marry
Derivation	Adj Adv	magnetic, magnetism simply, simple

EuroWordNet [5] is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). It uses the same structure as the English WordNet. EuroWordNet can solve cross-language problems, for example, words for different languages, such as English, French, Italian, German, are used to name the same entities.

3 Exploiting WordNet in Ontology Matching

Semantic similarity based on WordNet has been widely explored in Natural Language Processing and Information Retrieval. But most of these methods are applied in an ontology (e.g., WordNet). We will first show these methods, then we will discuss how to apply them in ontology matching.

Several methods for calculating semantic similarity between words in WordNets exist and can be classified into three categories:

- Edge-based methods: to measure the semantic similarity between two words is to measure the distance (the path linking) of the words and the position of the word in the taxonomy. That means the shorter the path from one node to another, the more similar they are (e.g., [27], [18], [24]).
- Information-based statistics methods: to solve the difficult problem to find a uniform link distance in edge-based methods, Resnik proposes an information-based statistic method [19].The basic idea is that the more information two concepts have in common, the more similar they are. This approach is independent of the corpus. For examples see [19], [13].
- Hybrid methods: combine the above methods, e.g., [21], [9], [4].

3.1 Edge-based Methods

Wu and Palmer [27] propose defining the similarity of two concepts based on the common concepts by using the path.

$$sim(C_1,C_2) = \frac{2*N_3}{N_1+N_2+2*N_3}, \tag{1}$$

where C_3 is the least common superconcept of C_1 and C_2. N_1 is the number of nodes on the path from C_1 to C_3. N_2 is the number of nodes on the path from C_2 to C_3. N_3 is the number of nodes on the path from C_3 to root.

Resnik [18] introduces a variant of the edge-based method, converting it from a distance to a similarity metric by subtracting the path length from the maximum possible path length.

$$sim_{edge}(w_1,w_2) = (2*MAX) - [min_{c_1,c_2} len(c_1,c_2)] \tag{2}$$

where $s(w_1)$ and $s(w_2)$ represent the set of concepts in the taxonomy that are senses of word w_1, w_2 respectively, c_1 overs $s(w_1)$, c_2 overs $s(w_2)$, MAX is the maximum depth of the taxonomy, and $len(c_1,c_2)$ is the length of the shortest path from c_1 to c_2.

Su defines the similarity of two concepts based on the distance of the two concepts in WordNet [24]. This can be done by finding the paths from one concept to the other concept and then selecting the shortest such path. Threshold like 11 is set for the top nodes of the noun taxonomy. That means not always a path can be found between two nouns. The WordNet similarity is used to adjust similarity value in his ontology matching system.

3.2 Information-based Statistics Methods

Resnik proposes an information-based statistic method [19]. First, it calculates the probability with concepts in the taxonomy, then follows information theory, the information content of a concept can be quantified as negative the log likelihood. Let $\mathscr{C}$ be set of concepts in the taxonomy. The similarity of two concepts is extent to the specific concept that subsumes them both in the taxonomy. Let the taxonomy be augmented with a function $p : \mathscr{C} \rightarrow [0,1]$, such that for any $c \in \mathscr{C}$, $p(c)$ is the probability of encountering concept c. If the taxonomy has a unique top node then its probability is 1. The information content of c can be quantified as $-\log p(c)$. Then

$$sim(c_1,c_2) = max_{c \in S(c_1,c_2)}[-\log p(c)], \tag{3}$$

where $S(c_1,c_2)$ is the set of concepts that subsume both c_1 and c_2. The word similarity (sim) is defined as

$$sim(w_1,w_2) = max_{c_1,c_2}[sim(c_1,c_2)], \tag{4}$$

where $s(w_1)$ and $s(w_2)$ represent the set of concepts in the taxonomy that are senses of word w_1, w_2 respectively, c_1 overs $s(w_1)$, c_2 overs $s(w_2)$.

Lin adapts Resnik's method and defines the similarity of two concepts as the ratio between the amount of information needed to state the commonality between them and the information needed to fully describe them [13].

$$sim(x_1,x_2) = \frac{2 \times \log p(c_0)}{\log p(c_1) + \log p(c_2)}, \tag{5}$$

where $x_1 \in c_1$ and $x_2 \in c_2$, c_0 is the most specific class that subsumes both c_1 and c_2. The ontology alignment tool RiMOM [28] includes Lin's approach in the system.

3.3 Hybrid Methods

Jiang and Conrath propose a combined model that is derived from the edge-based notion by adding the information content as a decision factor [9]. The information content $IC(c)$ of a concept c can be quantified as $-\log P(c)$. The link strength (LS) of an edge is the difference of the information content values between a child concept and its parent concept.

$$LS(c_i,p) = -\log(P(c_i|p)) = IC(c_i) - IC(p) \tag{6}$$

where child concept c_i is a subset of its parent concept p, . After considering other factors, e.g., local density, node depth, and link type, the distance function is:

$$Dist(w_1,w_2) = IC(c_1) + IC(c_2) - 2 \times IC(LSuper(c_1,c_2)), \tag{7}$$

where $LSuper(c_1,c_2)$ is the lowest super concept of c_1 and c_2.

Rodriguez presents another approach to determine similar entities based on WordNet. For example, it considers hypernym/hyponym, holonym/meronyms relations [21]. The similarity measure based on the normaliztion of Tversky's model and set theory functions (S) of intersection $|A \cap B|$ and difference $|A/B|$ is as follows:

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a,b)\,|A/B| + (1-\alpha(a,b))\,|B/A|} \tag{8}$$

where a and b are entity classes, A and B are the description sets of a and b (i.e., synonym sets, is-a or part-whole relations), α is a function that defines the relative importance of the non-common characteristics. For $is-a$ hierarchy, α is expressed in term of the depth of the entity classes.

$$\alpha(a,b) = \begin{cases} \dfrac{depth(a)}{depth(a)+depth(b)} & if\ depth(a) \leq depth(b) \\ 1 - \dfrac{depth(a)}{depth(a)+depth(b)} & if\ depth(a) > depth(b) \end{cases} \tag{9}$$

Petrakis et al. adapt Rodrigues approach and develop X-Similarity which relies on synsets and term description sets [4]. Equation 8 is replaced as plain set similarity (S) where A and B mean synsets or term description sets.

$$S(a,b) = max\frac{A \cap B}{A \cup B}, \tag{10}$$

The similarity between term neighborhoods $S_{neighborhoods}$ is computed per relationship type (e.g., Is-A and Part-Of) as

$$S_{neighborhoods}(a,b) = max\frac{A_i \cap B_i}{A_i \cup B_i}, \tag{11}$$

where i denote relation type. Finally,

$$Sim(a,b) = \begin{cases} 1,\ if\ S_{synsets}(a,b) > 0 \\ max(S_{neighborhoods}(a,b), S_{descriptions}(a,b))\ if\ S_{synsets}(a,b) = 0 \end{cases} \tag{12}$$

where $S_{descriptions}$ means the matching of term description sets. $S_{descriptions}$ and $S_{synsets}$ are calculated according equation 11.

Bath et al. adapt Jaro-Winkler (JW) metric to integrate WordNet or EuroWordNet in processing ontology matching [1]. Name similarity (NS) of two names N_1 and N_2

of two classes A and B (each name is a set of tokens, $N = \{n_i\}$) is defined as

$$NS'(N_1,N_2) = \frac{\sum_{n_1 \in N_1} MJW(n_1,N_2') + MJW(n_2,N_1')}{|N_1| + |N_2|} \tag{13}$$

where $MJW(n_i,N) = max_{n_j \in N} JW(n_i,n_j)$, $N_i' = N_i \cup \{n_k | \exists n_j \in N_i \bigcap n_k \in synset(n_j)\}$, $synset(n_j)$ is the set of synonyms of term n_j, $NS'(A,B) = NS'(N_1,N_2)$.

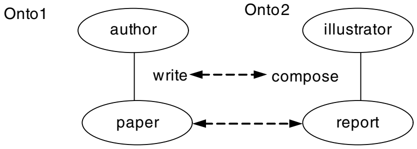
3.4 Applying WordNet Based Semantic Similarity Methods in Ontology Matching

Before applying the semantic similarity method in ontology matching, linguistic normalization is processed. Linguistic technologies transform each term to a standard form that can be easily recognized.

- Tokenisation consists of segmenting strings into sequences of tokens by a tokeniser which recognizes punctuation, cases, blank characters, digits, etc [6]. For example, $travel - agent$ becomes $< travel\ agent >$.
- Stemming is trying to remove certain surface marking words to root form. For example, words like *fishes* original form is *fish*.
- Stop-word [2] means that some words frequently appear in the text with lack of indexing consequence. Indexing is the process of associating one or more keywords with each document in information retrieval. For example, words like *the, this* and *of* in English, they appear often in sentences but have no value in indexing.
- Multiple part-of-speech. Each part-of-speech explains not what the word is, but how the word is used. In fact, the same word can be more than one part-of-speech (for instance, *backpacking* is both a noun and a verb in WordNet). When we compare the concept names which are made of single noun or noun phrase in the ontology, for these words it will be checked if they are nouns and if the answer is yes, we treat them as noun and disregard as verb [24].

WordNet based semantic similarity methods (see section 3.1, 3.2 and 3.3) can be used in two ways.

Fig. 2 Two simple ontologies.



- WordNet based semantic similarity methods can be applied to calculate entities similarities in two ontologies. For example, Figure 2 shows two simple ontologies Onto1 and Onto2. Property *write* in Onto1 and *compose* in Onto2 are synonyms in WordNet, we treat the labels of these two properties as equal even their string similarities are different. Since *paper* in Onto1 is the synonym of *report* in Onto2, they are treated as similar also.

There are two senses for the entry noun *author* hypernym relation in WordNet (version 2.1):

Sense 1

writer, author – (writes (books or stories or articles or the like) professionally (for money))

⟹ *communicator – (a person who communicates with others)*

⟹ *person, individual, someone, somebody, mortal, soul – (a human being; "there was too much for one person to do")*

⟹ *organism, being – (a living thing that has (or can develop) the ability to act or function independently)*

...

Sense 2

generator, source, author – (someone who originates or causes or initiates something; "he was the generator of several complaints")

⟹ *maker, shaper – (a person who makes things)*

⟹ *creator – (a person who grows or makes or invents things)*

⟹ *person, individual, someone, somebody, mortal, soul – (a human being; "there was too much for one person to do")*

...

There is one sense for the entry noun *illustrator* hypernym relation in WordNet (version 2.1):

illustrator – (an artist who makes illustrations (for books or magazines or advertisements etc.))

⟹ *artist, creative person – (a person whose creative work shows sensitivity and imagination)*

⟹ *creator – (a person who grows or makes or invents things)*

⟹ *person, individual, someone, somebody, mortal, soul – (a human being; "there was too much for one person to do")*

...

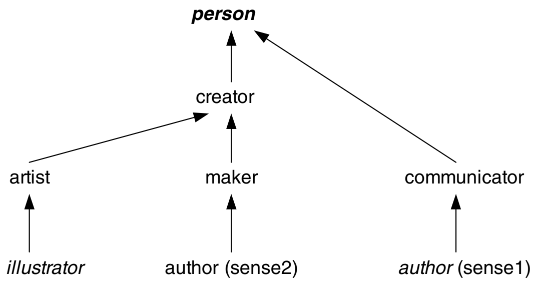


Fig. 3 The fragment of noun senses with *author* and *illustrator* in WordNet taxonomy.

Figure 3 presents the fragment of nouns with *author* and *illustrator* in WordNet taxonomy. If *author*is used in Onto1 and *illustrator* is used in Onto2 (see Figure 2), they have the common superconcept (hypernym) *person* in WordNet (see Figure 3), and we can apply WordNet based semantic similarity methods (see section 3.1, 3.2 and 3.3) to get similarity between *illustrator* and *author*.

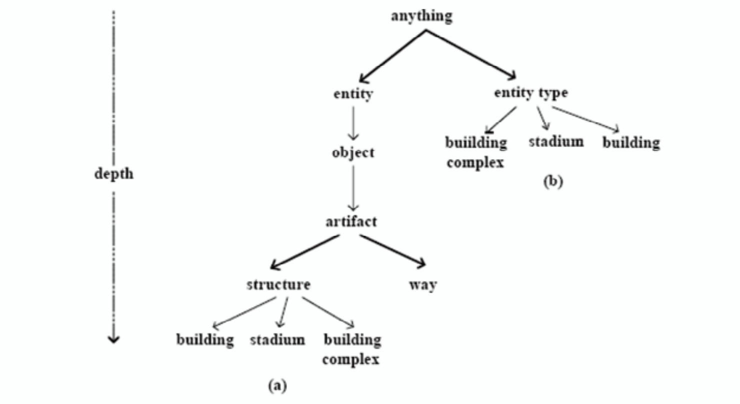


Fig. 4 Connecting independent ontologies: (a) partial WordNet ontology and (b) partial SDTS ontology. Source: [21]

- Rodriguez method [21] and X-Similarity [4] are independent from WordNet. They can be applied in ontology matching directly as structure similarity method if two independent ontologies have a common superconcept. For example, Figure 4 (see source [21]) shows two independent ontologies, *anything* is their common superconcept. Based on string similarity results, the structure similarity (e.g., similarity between *building*^{w} in WordNet and *building*^{s} in SDTS) can be calculated through Rodriguez method [21] and X-Similarity [4].

3.5 Evaluation of Semantic Similarity Methods

WordNet-Similarity [26] has implemented several WordNet-based similarity measures, such as Leacock-Chodorow [10], Jiang-Conrath [9], Resnik [18], Lin [13], Hirst-St-Onge [8], Wu-Palmer [27], Banerjee-Pedersen [15], and Patwardhan [15] in a Perl package.

Petrakis et al. [4] implement a "Semantic Similarity System" and evaluate several semantic similarity measures: Rada [17], Wu-Palmer [27], Li [12], Leacock-Chodorow [10], Richardson [20], Resnik [19], Lin [13], Lord [16], Jiang-Conrath [9], X-Similarity [4], Rodriguez [21]. Their evaluation in the same ontology is based on Miller and Charles [14] with the human relevance results. The higher the correlation of a method, the better the method is (i.e., the closer it is to the results of human judgement). They also evaluate Rodriguez [21] and X-Similarity [4] methods in different ontologies (ontology matching).

SimPack [23] implements methods such as Jiang-Conrath [9], Lin [13], Resnik [19]. These methods have been evaluate by Budanitsky and Hirst [3].

Table 1 compares different WordNet-based similarity measures in WordNet-Similarity, Semantic Similarity System and SimPack:

Table 1 Implemented WordNet-based similarity measures in WordNet-Similarity, Semantic Similarity System and SimPack

WordNet-Similarity	Semantic Similarity System	SimPack
Leacock-Chodorow [10]	Leacock-Chodorow [10]	
Jiang-Conrath [9]	Jiang-Conrath [9]	Jiang-Conrath [9]
Resnik [18]	Resnik [18]	Resnik [18]
Lin [13]	Lin [13]	Lin [13]
Hirst-St-Onge [8]		
Wu-Palmer [27]	Wu-Palmer [27]	
Banerjee-Pedersen [15]		
Patwardhan [15]		
	Rada [17]	
	Li [12]	
	Richardson [20]	
	Lord [16]	
	X-Similarity [4]	
	Rodriguez [21]	

4 Conclusions

In this paper, we present different WordNet-based semantic similarity measures from edge-based methods to information-based statistic methods and their hybrid methods. We also discuss how to apply them in the ontology matching. Finally, we show several tools that implemented the semantic similarity measures and their evaluation.

Acknowledgements Part of this work was financed by the Hamrin Foundation (Hamrin Stiftelsen), project Media Information Logistics. Special thanks to Lars Ahrenberg for valuable comments and suggestions on this paper.

References

1. Bach, T.L., Dieng-Kuntz, R., Gandon, F.: On ontology matching problems (for building a corporate semantic web in a multi-communities organization). In: Proc. 6th International

Conference on Enterprise Information Systems (ICEIS), pp. 236–243. Porto (PT) (2004)

2. Belew, R.K.: Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW. Cambridge University Press (2001)
3. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006)
4. Euripides G.M. Petrakis Giannis Varelas, A.H.P.R.: Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies. In: In 4th Workshop on Multimedia Semantics (WMS'06), pp. 44–52 (2006)
5. EuroWordNet: http://www.illc.uva.nl/eurowordnet
6. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag (2007)
7. Ferrer-i-Cancho, R.: The structure of syntactic dependency networks: insights from recent advances in network theory. In: L. V., A. G. (eds.) Problems of quantitative linguistics, pp. 60–75 (2005)
8. Hirst, G., St-Onge, D.: Lexical chains as representation of context for the detection and correction malapropisms (1997)
9. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy (1997)
10. Leacock, C., Chodorow, M.: Combining local context and wordnet similarity for word sense identification. An Electronic Lexical Database pp. 265–283 (1998)
11. Leacock, C., Miller, G.A., Chodorow, M.: Using corpus statistics and wordnet relations for sense identification. Comput. Linguist. **24**(1), 147–165 (1998)
12. Li Y.; Bandar, Z.M.D.: An approach for measuring semantic similarity between words using multiple information sources. Transactions on Knowledge and Data Engineering **15**(4), 871–882 (July-Aug. 2003). DOI 10.1109/TKDE.2003.1209005
13. Lin, D.: An information-theoretic definition of similarity. In: Proc. 15th International Conf. on Machine Learning, pp. 296–304. Morgan Kaufmann, San Francisco, CA (1998)
14. Miller George A., C.W.G.: Contextual correlates of semantic similarity. Language and Cognitive Processes **6**, 1–28 (1991)
15. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pp. 241–257. Mexico City, Mexico (2003)
16. P.W.Lord R.D. Stevens, A.B., C.A.Goble: Investigating semantic similarity measures across the gene ontology: (2002)
17. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on **19**(1), 17–30 (1989)
18. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: IJCAI, pp. 448–453 (1995)
19. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research **11**, 95–130 (1999)
20. Richardson, R., Smeaton, A.F., Murphy, J.: Using WordNet as a knowledge base for measuring semantic similarity between words. Tech. Rep. CA-1294, Dublin, Ireland (1994)
21. Rodriguez, M., Egenhofer, M.: Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering **15**(2), 442–456 (2003)
22. SemanticWeb: http://www.semanticweb.org/
23. SimPack: http://www.ifi.unizh.ch/ddis/simpack.html
24. Su, X.: Semantic enrichment for ontology mapping. Ph.D. thesis, Dept. of Computer and Information Science, Norwegian University of Science and Technology (2004)
25. WordNet: http://wordnet.princeton.edu
26. WordNet-Similarity: http://www.d.umn.edu/~tpederse/similarity.html
27. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 133 –138. New Mexico State University, Las Cruces, New Mexico (1994)
28. Yi Li Jie Tang, D.Z.J.L.: Toward strategy selection for ontology alignment. In: Proceedings of the 4th European Semantic Web Conference 2007 (ESWC2007) (2007)