Harnessing different knowledge sources to measure semantic relatedness under a uniform model

Ziqi Zhang **Anna Lisa Gentile** **Fabio Ciravegna**
Department of Computer Science, University of Sheffield
211 Portobello, Regent Court
Sheffield, S1 4DP
z.zhang@dcs.shef.ac.uk a.l.gentile@dcs.shef.ac.uk f.ciravegna@dcs.shef.ac.uk

Abstract

Measuring semantic relatedness between words or concepts is a crucial process to many Natural Language Processing tasks. Exiting methods exploit semantic evidence from a single knowledge source, and are predominantly evaluated only in the general domain. This paper introduces a method of harnessing different knowledge sources under a uniform model for measuring semantic relatedness between words or concepts. Using Wikipedia and WordNet as examples, and evaluated in both the general and biomedical domains, it successfully combines strengths from both knowledge sources and outperforms state-of-the-art on many datasets.

1 Introduction

Semantic relatedness (SR) measures how much two (strings of) words or concepts are related by encompassing all kinds of relations between them (Strube and Ponzetto, 2006). It is more general than semantic similarity. SR is often an important pre-processing step to many complex Natural Language Processing (NLP) tasks, such as Word Sense Disambiguation (Leacock and Chodorow, 1998; Han and Zhao, 2010), and information retrieval (Finkelstein et al., 2002). In the biomedical domain, SR is an important technique for discovering gene functions and interactions (Wu et al., 2005; Ye et al., 2005).

There is an abundant literature on measuring SR between words or concepts. Typically, these methods extract semantic evidence of words and concepts from a background knowledge source, with which their relatedness is assessed. The knowledge sources can be unstructured documents or (semi-)structured resources such as Wikipedia, WordNet, and domain specific ontologies (e.g., the Gene Ontology[1]).

In this paper, we identify two issues that have not been addressed in the previous works. First, existing works typically employ a single knowledge source of semantic evidence. Research (Strube and Ponzetto, 2006; Zesch and Gurevych, 2010; Zhang et al., 2010) has shown that the accuracy of an SR method differs depending on the choice of the knowledge sources, and there is no conclusion which knowledge source is superior to others. Zhang et al. (2010) argue that this indicates different knowledge sources may complement each other. Second, the majority of SR methods have been evaluated in general domains only, except a few earlier WordNet-based methods that have been adapted to biomedical ontologies and evaluated in that domain (Lord et al., 2003; Pedersen et al., 2006; Pozo et al., 2008). Given the significant attention that SR has received in specific domains (Pesquita et al., 2007), evaluation of SR methods in specific domains is increasingly important.

This paper addresses these issues by proposing a generic and uniform model for computing SR between words or concepts using multiple knowledge sources, and evaluating the proposed method in both general and specific domains. The method combines and integrates semantic evidence of words or concepts extracted from any knowledge source in a generic graph representation, with which the SR between concepts or words is computed. Using two of the most popular general-domain knowledge sources,

[1] http://www.geneontology.org/, last retrieved in Mar. 2011

Wikipedia and WordNet as examples, the method is evaluated on 7 benchmarking datasets, including three datasets from the biomedical domain and four from the general domain. It has achieved excellent results: compared to the baselines that use each single knowledge sources, combining both knowledge sources has improved the accuracy on all datasets by 2~11%; compared to state-of-the-art on the general domain datasets, the method achieves the best results on three datasets; and on the other three biomedical datasets, it obtains the best result in one case; and second and third best results on the other two among eight participating methods, where all other competitors exploit some domain-specific knowledge sources.

The remainder of this paper is organized as follows. Section 2 discusses related work; Section 3 presents the proposed method; Section 4 describes the experiments and evaluation; Section 5 discusses results and findings; Section 6 concludes this paper.

2 Related work

2.1 SR methods

Methods for computing SR can be classified into *path based*, *Information Content (IC) based*, *statistical* and *hybrid methods*. *Path based* methods (Hirst and St-Onge, 1998; Leacock and Chodorow, 1998; Pekar and Staab, 2002; Rada et al., 1989; Wu and Palmer, 1994) measure SR between words or concepts as a function of their distance in a semantic network, usually calculated based on the path connecting the words or concepts by certain semantic (typically *is-a*) links. *IC based methods* (Jiang and Conrath, 1997; Lin, 1998; Pirro et al., 2009; Resnik, 1995; Seco et al., 2004) assess relatedness between words or concepts by the amount of information they share, usually determined by a higher level concept that subsumes both concepts in a taxonomic structure. *Statistical methods* measure relatedness between words or concepts based on their distribution of contextual evidence. This can be formalized as co-occurrence statistics collected from unstructured documents (Chen et al., 2006; Cilibrasi and Vitanyi, 2007; Matsuo et al., 2006), or distributional concept or word vectors with features extracted from either unstructured documents (Harrington, 2010; Wojtinnek and Pulman, 2011) or (semi-)structured knowledge resources (Agirre et al., 2009; Gabrilovich and Markovitch, 2007; Gouws et al., 2010; Zesch and Gurevych, 2007; Zhang et al., 2010). *Hybrid methods* combine different purebred methods in certain ways. For example Riensche et al. (2007) employ both an *IC based* method (Resnik, 1995) and a *statistical* method (cosine vector similarity) in their study. Pozo et al. (2008) derive a taxonomy of terms from unstructured documents by applying hierarchical clustering based on corpus statistics, then apply *path based* method on this taxonomy to compute SR. Han and Zhao (2010) use one *IC based* method and two *statistical* methods to compute SR, then derive an aggregated score.

2.2 SR knowledge sources and domains

Computing SR requires background knowledge about concepts or words, which can be extracted from unstructured corpora, semi-structured and structured knowledge resources. Unstructured corpora are easier to create and cheaper to maintain, however, semantic relations between words or concepts are implicit. Methods (Chen et al., 2006; Cilibrasi and Vitanyi 2007; Matsuo et al., 2006) that exploit unstructured corpora typically depend on distributional statistics, and thus may ignore important semantic evidences present in (semi-)structured knowledge sources (Pan and Farrell, 2007). Recent studies (Harrington, 2010; Pozo et al., 2008; Wojtinnek and Pulman, 2011) propose to pre-process a corpus to learn a semantic network, with which SR is computed. This creates high pre-processing cost; also, the choice of corpus and its size often have a direct correlation with the accuracy of SR methods (Batet et al., 2010).

(Semi-)Structured knowledge sources on the other hand, organize semantic knowledge about concepts and words explicitly and interlink them with semantic relations. They have been popular choices in the studies of SR, and they include lexical resources such as WordNet, Wiktionary, and (semi-)structured encyclopedic resources such as Wikipedia. WordNet has been used in earlier studies (Hirst and St-Onge, 1998; Jiang and Conrath, 1997; Lin, 1998; Leacock and Chodorow 1998; Resnik, 1995; Seco et al., 2004; Wu and Palmer, 1994) and is still a preferred knowledge source in recent works (Agirre et al., 2009). However, its effectiveness may be hindered by its lack of coverage of specialized lexicons and domain specific concepts (Strube and Ponzetto,

2006; Zhang et al., 2010). Wikipedia and Wiktionary are collaboratively maintained knowledge sources and therefore may overcome this limitation. Wikipedia in particular, is found to have reasonable coverage of many domains (Holloway et al., 2007; Halavais, 2008). It has become increasingly popular in SR studies recently. However, research (Zesch and Gurevych, 2010) have shown that methods based on Wikipedia have no clear advantage over WordNet-based methods on some general domain datasets in terms of accuracy, while Zhang et al. (2010) argue that different knowledge sources may complement each other, and SR methods may benefit from harnessing different knowledge sources.

Several studies (Lord et al., 2003; Pedersen et al., 2006; Petrakis et al., 2006; Pozo et al., 2008) have adapted state-of-the-art to domain specific knowledge sources (e.g., the Gene Ontology, the MeSH[2]) and evaluated them therein. Despite these efforts, a large proportion of state-of-the-art is still only evaluated in the general domain.

2.3 SR methods similar to this work

Few works have attempted at combining different knowledge sources in SR studies, especially (semi-)structured knowledge sources. The closest studies are Han and Zhao (2010) and Tsang and Stevenson (2010). Han and Zhao firstly compute SR between words using three state-of-the-art SR methods separately. Next, one score is chosen subject to an arbitrary preference order, and used to create a connected graph of weighted edges between words. A recursive function is then applied to the graph to compute final SR scores between words. Essentially, each SR method is applied in isolation and features from different sources are used separately with each distinctive method. Although this retains advantages of each method, the limitations of them are also combined.

Tsang and Stevenson (2010) combine WordNet and unstructured documents by weighing each word found in WordNet using its frequency observed in a large corpus. The frequencies however, are sensitive to the choice of corpus, thus different corpora may result in different accuracies. Furthermore, their method is only applicable to computing SR between pairs of sets of words or concepts.

3 Methodology

We define a set of requirements for SR methods that harness different knowledge sources:

- It should improve over the same method based on a single knowledge source
- It should be generic and applicable to any knowledge source
- It should be robust in dealing with knowledge source specific features but also tolerate the quality and coverage issues of individual knowledge source

Our method of harnessing different knowledge sources contains four steps. Firstly (Section 3.1), each word or word segment is searched in each knowledge source to identify their *contexts* that is specific to that knowledge source. We define a *context* as the representation of meaning or a concept for a word. In the following, we say that each *context* is associated with a distinct concept. Secondly (Section 3.2), for each concept of an input word, features are extracted from its *context* and a graph representation of each concept and their features is created. Thirdly (Section 3.3), cross-source *contexts* are mapped where they refer to the same concept, thus their features from different sources can be combined to derive an enriched representation. This creates a final, uniform graph representation where input words are connected by shared features of their underlying candidate concepts. Then (Section 3.4) the graph is submitted to a generic algorithm to compute SR between words.

In the following, we discuss details with respect to different types of knowledge sources, while focusing on Wikipedia and WordNet in our experiments for two reasons. First, they are used by the majority of SR methods and are therefore most representative knowledge sources. Second, they have strongly distinctive and complementary characteristics, which make ideal testbeds for the requirements. On one hand, WordNet is a lexical resource containing rich and strict semantic relations between words, but lacks coverage of specialized vocabularies. On the other hand, Wikipedia is a semi-structured resource with good coverage of domains and named entities, but the semantic knowledge is organized in a looser way.

[2] http://www.nlm.nih.gov/mesh/ last retrieved in March 2011

3.1 Context retrieval

Given a pair of words or word segments, we firstly identify *contexts* representing the underlying meanings or concepts from each knowledge source. For lexical resources, this could be distinctive word senses. In **WordNet (WN)**, a context corresponds to a single synset, which corresponds to a concept. We search each word in WordNet and extract all possible synsets. Let *w* be a word or word segment (e.g., "cat"), and $C_w^{wn} = \{c_{1_w}^{wn}, c_{2_w}^{wn} \dots c_{k_w}^{wn}\}$ be the set of *k* concepts of *w* extracted from WordNet.

Using **Wikipedia (WK)** as an example semi-structured resource, the *context* can be an article that describes a unique concept. Thus we search for underlying articles that describe different concepts. Firstly, we search *w* in Wikipedia, where three situations may be anticipated. If a single non-disambiguation page describing a concept is returned, the concept is selected and the retrieval is complete. In the second case, a disambiguation page linking to all possible concept pages may be returned. This page lists all underlying concepts and entities referenced by *w* as links and a short description with each link. In this case, we always keep the first concept page, which is found often to be the most common sense of the word; additionally, we select other concept pages whose short descriptions contain the word *w*. We do not select all linked pages because many of these in fact link to a concept relevant to *w*, but not necessarily a candidate sense of *w*. Thirdly, if no pages are returned for *w*, we search for the most relevant page using *w* as keyword(s) in an inverted index of all Wikipedia pages (e.g., via search engines). We denote concepts retrieved from Wikipedia as $C_w^{wk} = \{c_{1_w}^{wk}, c_{2_w}^{wk} \dots c_{k_w}^{wk}\}$.

For unstructured sources such as documents, a simple approach could be defining a word *context* as a text passage around each occurrence of *w*, and grouping similar *contexts* of *w* as representation of its underlying meanings, or concepts. Alternatively, more complex approaches such as Pozo et al. (2008) and Harrington (2010) may be applied to extract a lexical network of words, whereby similar methods to WordNet can be applied.

3.2 Feature extraction and representation

Next, for each concept identified from a knowledge source, features are extracted from their corresponding *contexts*. In our case, for each $c \in C_w^{wk}$, we follow the work by Zhang et al. (2010) to extract four types of features from their corresponding Wikipedia pages. Figure 1 shows an example representation of a concept and its Wikipedia features:

- Words from page titles and redirection links (can be considered as synonyms)
- Words from categories, used as higher level hypernyms in some studies (Zesch et al., 2010; Strube and Ponzetto, 2006)
- Words from outgoing links
- Top *n* most frequent words from a page

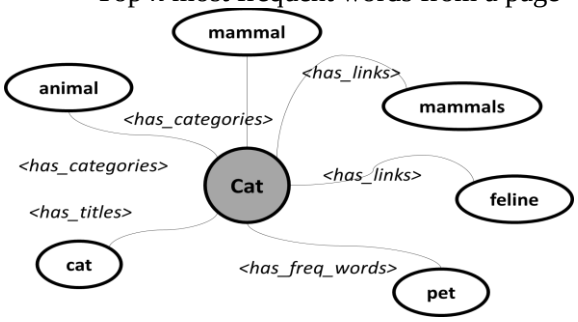

Figure 1. Representation of the concept "cat, the mammal" using different types of features extracted from Wikipedia. The shaded circle represents the concept; ovals represent feature values; edges connecting feature values to the concept and <labels> represent feature types

For each $c \in C_w^{wn}$, we extract ten features from WordNet: hypernyms, hyponyms, meronyms, holonyms, synonyms, antonyms, attributes, "see also" words, "related" words, and gloss. These are also represented in the same way as in Figure 1.

With unstructured sources, contextual words can be used as features. Alternatively, if a lexical network is extracted, features may be extracted in a similar way to those of WordNet.

Additionally, with WordNet and Wikipedia, we also propose several intra-resource feature merging strategies to study the effect of **feature diversification**. This is because, while some approaches (such as Agirre et al., 2009; Harrington, 2010; Yeh et al., 2009) do not distinguish different feature types in graph construction, or adopt a bag-of-words feature representation (such as Zesch and Gurevych, 2010), others (such as Yazdani and Popescu-Belis, 2010; Zhang et al., 2010) have used differentiated

feature types and weights in their model. We therefore carry out studies to investigate this issue. Specifically, for the original four Wikipedia features, we create a bag-of-words feature that simply merges all feature types (i.e., all edges in Figure 1 will have the same label). For the original ten WordNet features, we propose two merged representations corresponding to that of Wikipedia, so as to support the studies of feature enrichment in the following section. We introduce a bag-of-words feature that collapses all different feature types, and a four-feature representation as follow:

- *wn-synant* merges WordNet synonyms and antonyms.
- *wn-hypoer* merges WordNet hypernyms and hyponyms, collectively representing features by "*is-a*" semantic relation
- *wn-assc* merges WordNet meronyms, holonyms, related and "see also", which are features corresponding to associative relations
- *wn-dist* merges WordNet gloss and attributes that generally describe a concept.

3.3 Concept mapping and feature enrichment

Our method essentially harnesses different knowledge sources by combining features extracted from different sources in a uniform model. This requires two sub-processes: **cross-source concept mapping** and **cross-source feature enrichment**.

In **cross-source concept mapping,** concepts extracted from different knowledge sources are mapped according to similar meanings such that cross-source features can be combined. To do so, we select the concepts from one knowledge source as the reference concept set; then concepts from other knowledge sources are mapped to reference concepts of similar meanings. There can be different criteria of choosing reference knowledge source concepts. Empirically, we found it necessary to choose the knowledge source with broader coverage and richer features. This will be discussed later in Section 5. Following this strategy, in our example, C_w^{wk} is chosen as reference concepts, and for each $c_w^{wk} \in C_w^{wk}$ we select a $c_w^{wn} \in C_w^{wn}$ such that c_w^{wk} and c_w^{wn} refer to the same meaning. To do so, we apply a simple maximum set overlap metric to their feature values. Let *F(c)* be a function that returns all feature values of *c* as bag-of-words, then for each $c_w^{wk} \in C_w^{wk}$, it is mapped to a c_w^{wn} such that $|F(c_w^{wn})| \cap |F(c_w^{wk})|$ is maximized among all $c_w^{wn} \in C_w^{wn}$. The resulting concept candidates are denoted as C_w^{mapd}, where $c_w^{mapd} = \{c_w^{wk}, c_w^{wn}\}$ is a mapped set of concepts potentially referring to the same meaning. If $C_w^{wn} = \emptyset$ then $C_w^{mapd} = C_w^{ref}, i.e., C_w^{wk}$.

Next, **cross-source feature enrichment** creates a uniform feature representation for each mapped sets of concepts. The process can be considered as enriching the features from one knowledge source with others. The most straightforward approach is to simply collect features extracted from each knowledge source on to a single graph, retaining the diversity in feature types. For example, Figure 2 shows a graph representation based on the collection of the four Wikipedia features and the four derived WordNet features. We refer to this approach as "***feature combination***".

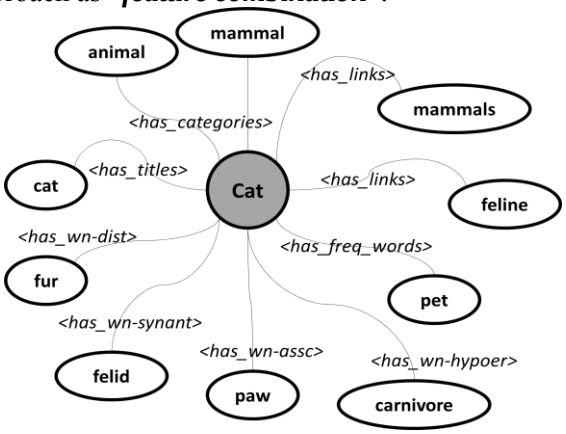

Figure 2. Representation of "cat, the mammal" after ***concept mapping*** and ***feature combination***

On the other hand, cross-source features may be merged according to their semantics. For example, WordNet and Wikipedia contain features based on synonyms of concepts; while Wikipedia and unstructured documents contain word distributional features. Thus we define "***feature integration***" as merging feature types from different knowledge sources into single types of features based on their similarity in semantics. With WordNet and Wikipedia, we integrate features as below (Figure 3):

- *merged-synant* merges Wikipedia page titles and redirection links with *wn-synant*
- *merged-hypoer* merges merges Wikipedia categories with *wn-hypoer*

- *merged-assc* merges Wikipedia links with *wn-assc*. We consider Wikipedia links bear other associative relations and are therefore merged with features extracted by other WordNet relations
- *merged-dist* merges Wikipedia frequent *n* words with *wn-dist*.

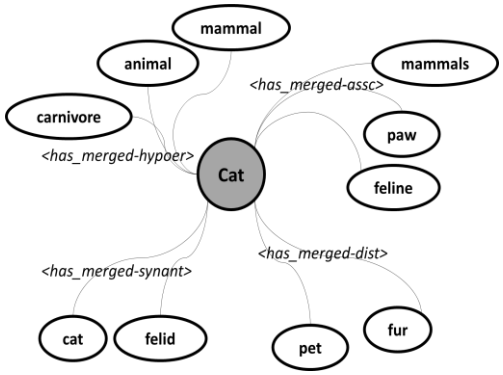

Figure 3. Representation of "cat, the mammal" after ***concept mapping*** and ***feature integration***

Note that the difference between cross-source ***feature combination*** and ***integration*** is that the former introduces more *types* of features, whereas the latter retains same number of feature types but increases *feature values* for each type. Both have the effect of establishing additional path (via features) between concepts, but in different ways.

With intra-resource **feature diversification,** cross-source ***feature combination*** and ***feature integration***, we create a total of nine intra- and cross-source feature representations to be tested with the uniform random walk model:

- four types of Wikipedia features (*wk-4F*)
- one type of Wikipedia features (*wk-1F*)
- ten types of WordNet features (*wn-10F*)
- four types of WordNet features (*wn-4F*)
- one type of WordNet features (*wn-1F*)
- *wk-4F combines wn-4F: wk-4F+wn4F,C*
- *wk-4F integrates wn-4F: wk-4F+wn4F,I*
- *wk-1F combines wn-1F: wk-1F+wn1F,C*
- *wk-1F integrates wn-1F: wk-1F+wn1F,I*

3.4 Computing SR using the graph

The algorithm for computing SR using the graph is based on the idea of random walk. It formalizes the idea that taking successive steps along the paths in a graph, the "easier" it is to arrive at a target node starting from a source node, the more related the two nodes are. Following the previous steps, the feature representations of all candidate concepts relevant to the input word pairs are joined, which creates a single undirected, weighted, bi-partite graph. Let *G = (V, E)* be the graph, where *V* is the set of nodes (concepts and feature values); *E* is the set of edges (feature types) that connect concepts and features. As shown in Figure 4, different concepts are connected if they share same values of same types of features, namely, there exists a path that connects one concept to another.

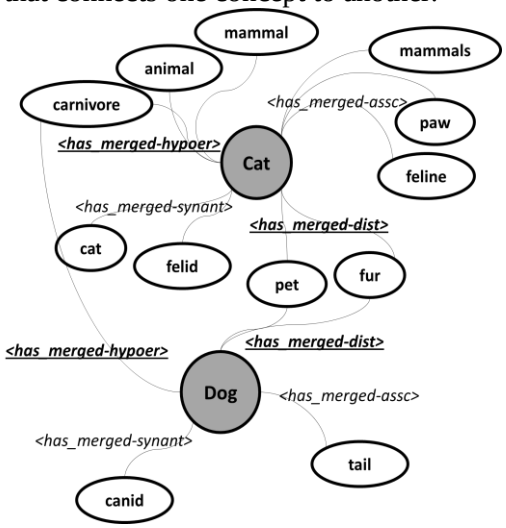

Figure 4. Paths are established between different concepts if they share values of same feature types <**bold underlined**>

Using Figure 4 it is easier to comprehend the difference between ***feature combination*** and ***integration***. Since concept nodes can only be connected by same types of edges (feature types), ***feature combination*** increases the chances of connectivity by adding in more types of edges, while ***integration*** merges similar types of edges across knowledge sources and increases the number of feature nodes connected by each type.

From the graph, we start by building an adjacency matrix *W* of initial probability distribution:

$$W_{ij} = \begin{cases} \sum_{l_k \in L} \frac{w(l_k)}{|(i,\cdot) \in E : l(i,\cdot) = l_k|}, (i,j) \in E \\ 0, otherwise \end{cases} \quad [1]$$

Where W_{ij} is the i^{th}-line and j^{th}-column entry of *W*, indexed by *V; l(i, j)* is a function that returns the type of edge (i.e., type of feature) connecting nodes *i* and *j*; *L* is the set of all possible types; *w(l)* returns the weight for that type. Essentially, *L* is the collection of all feature types, and *w(l)* assigns

a weight to a particular feature type. Next, we compute the transition probability matrix $P^{(t)}(j|i) = [(D^{-1}W)^t]_{ij}$ ($D_{ii} = \sum_k W_{ik}$), which returns the probability of reaching other nodes from a starting node on the graph after *t* steps. In this method, we follow the work by Rowe and Ciravegna (2010) to set *t=2* in order to preserve locally connected nodes. Next, we extract the probability vectors corresponding to concept nodes from *P*, and compute pair-wise relatedness using the cosine function. Effectively, this formalizes the notion that two concepts related to a third concept is also semantically related, which is similar to the hypothesis proposed by Patwardhan and Pedersen (2006) in their method based on second-order context vectors. The final SR between the input word pair is the maximum pair-wise concept SR.

4 Experiment and evaluation

We evaluate the method based on correlation against human judgment (gold standard) on seven benchmarking datasets covering both general and technical domains. These include four general domain datasets: the Rubenstein and Goodenough (1965) dataset containing 65 pairs of nouns (RG65); the Miller and Charles (1991) dataset that is a subset of the RG-65 dataset and contains 30 pairs (MC30); the Finkelstein et al. (2002) dataset with 353 pairs of words, including nouns, verbs, adjectives, as well as named entities. This contains two subsets, a set of 153 pairs (Fin153) and a set of 200 (Fin200) pairs each annotated by a different groups of annotators. Zesch and Gurevych (2010) show largely varying Inter-Annotator-Agreement (IAA) between the two sets (Table 1), and argue that they should be treated as separate datasets. Three biomedical datasets are selected to evaluate domain-specific performance of the proposed method. These include a set of 36 MeSH term pairs in Petrakis et al. (2006) (MeSH36), 30 pairs of medical terms annotated by a group of physicians as in Pedersen et al. (2006) (Ped30-p) and the same set annotated by a different group of medical coders (Ped30-c). Table 1 shows statistics of the seven datasets.

The correlation is computed using the Spearman rank order coefficient for two reasons. First, it is a better metric than other alternatives (Zesch and Gurevych, 2010). Second, it is consistent with the majority of studies such that results can be compared.

Dataset	Size	Domain	IAA
MC30	30	General	0.9
RG65	65	General	0.8
Fin153	153	General	0.73
Fin200	200	General	0.55
Ped30-p	30	Biomedical	0.68
Ped30-c	30	Biomedical	0.78
MeSH36	36	Biomedical	-

Table 1: Information of benchmarking datasets

We distribute feature weights *w(l)* across different feature types *L* evenly in each feature representation. Although Zhang et al. (2010) show that discriminated feature weights leads to improved accuracy; this is not the focus of this study. Since we aim to investigate the effects of harnessing different knowledge sources, we obtained baseline performances by applying the method to those feature representations based on single knowledge sources (i.e., *wk-4F, wk-1F, wn-10F, wn-4F, wn-1F*). Tables 2 and 3 show the best results obtained with baselines and corresponding knowledge sources and feature representation.

Dataset	Corr.	Feature	Coverage (% pairs)
MC30	0.77	*wn-1F*	77%
RG65	0.71	*wn-1F*	65%
Fin153	0.45	*wn-4F*	82%
Fin200	0.35	*wn-4F*	76%
Ped30-p	0.66	*wn-4F*	33%
Ped30-c	0.8	*wn-4F*	33%
MeSH36	0.49	*wn-1F*	50%

Table 2: Correlation obtained using *WordNet*. Many word pairs are not covered due to sparse feature space and lack of coverage. **Only covered pairs are accounted**.

Dataset	Corr.	Feature
MC30	0.74	*wk-1F*
RG65	0.67	*wk-1F*
Fin153	0.7	*wk-1F*
Fin200	0.51	*wk-4F*
Ped30-p	0.53	*wk-4F*
Ped30-c	0.58	*wk-4F*
MeSH36	0.73	*wk-4F*

Table 3: Correlation obtained using only Wikipedia. All word pairs are 100% covered.

Tables 4 – 6 show results obtained with enriched feature representation.

	Combination (**C**)		Integration (**I**)	
Dataset	*wn-4F + wk-4F*	*wn-1F + wk-1F*	*wn-4F + wk-4F*	*wn-1F + wk-1F*
MC30	0.77	0.8	0.8	0.79
RG65	0.74	0.73	0.73	0.729
Fin153	0.73	0.75	0.74	0.73
Fin200	0.52	0.54	0.53	0.54
Ped30-p	0.63	0.52	0.64	0.47
Ped30-c	0.64	0.52	0.67	0.49
MeSH36	0.7	0.694	0.75	0.7

Table 4: Correlation obtained using both knowledge sources. Word pairs are 100% covered.

	KS and # of feature types			
	WN	WK	WK+WN,**C**	WK+WN, **I**
MC30	1	1	1	4
RG65	1	1	4	4
Fin153	4	1	1	4
Fin200	4	4	1	1
Ped30-p	4	4	4	4
Ped30-c	4	4	4	4
MeSH36	1	4	4	4

Table 5: Number of feature types with which best results are obtained on each dataset. **KS**: Knowledge Source

	Single KS	Multiple KS		Impr.
Dataset	Best corr.	Best corr.	Strategy	
MC30	0.74	0.8	C/I	0.06
RG65	0.67	0.74	C	0.07
Fin153	0.7	0.75	C	0.05
Fin200	0.51	0.54	C/I	0.03
Ped30-p	0.53	0.64	I	0.11
Ped30-c	0.58	0.67	I	0.09
MeSH36	0.73	0.75	I	0.02

Table 6: Improvement achieved by harnessing multiple KSs. Best correlation with single KS is based on Wikipedia, which provides 100% coverage of word pairs.

Tables 7 and 8 compare our method against state-of-the-art. For Table 8, figures for other state-of-the-art systems can be found in corresponding publications; while we only list the best performing systems for comparison.

	MC30	RG65	Fin153	Fin200	KS
best of WN+WK	**0.8**	0.74	**0.75**	**0.54**	Both
Rad89*	0.75	0.79	0.33	0.24	WN
LC98*	0.75	0.79	0.33	0.24	WN
WP94*	0.77	0.78	0.38	0.24	WN
HS98*	0.76	0.79	0.33	0.32	WN
Res95*	0.72	0.74	0.35	0.26	WN
JC97*	0.68	0.58	0.28	0.10	WN
Lin98*	0.67	0.60	0.27	0.17	WN
Zes07*	0.77	0.**82**	0.6	0.51	WK
GM07*	0.67	0.75	0.69	0.51	WK
Zha10	0.71	0.76	0.71	0.46	WK

Table 7[3]: Comparison against state-of-the-art in the general domain. (* figures from Zesch and Gurevych, 2010)

	Ped30-p	Ped30-c	MeSH36	KS
best of WN+WK	0.64	0.67	**0.75**	WN+ WK
Pet06 best	-	-	0.74	MeSH
Ped06 best	**0.84**	**0.75**	-	GO, D
Ped06 second	0.62	0.68	-	GO, D

Table 8[4]: Comparison against state-of-the-art in the biomedical domain. GO – Gene Ontology; D – document sets.

Given the fact that some datasets (i.e., MC30, Ped30-p, Ped30-c, MeSH36) have a relatively low sample size, we cannot always be sure that correlation values are accurate or occurred by chance. Therefore, we measure the statistical significance of correlation by computing the *p-value* for the correlation values reported for our system in Tables 7 and 8. For all cases, a *p-value* of less than 0.001 is obtained, which indicates that correlation values are statistically significant.

[3] Rada (1989) (Rad89); Leacock and Chodorow (1998) (LC98); Wu and Palmer (1994) (WP04); Hirst and St-Onge (1998) (HS98); Resnik (1995) (Res95); Jiang and Conrath (1997) (JC97); Lin (1998) (Lin98); Zesch and Gurevych (2007) (ZG07); Gabrilovich and Markovitch (2007) (GM07); Zhang et al. (2010) (Zha10)

[4] Petrakis et al. (2006) (Pet06); Pedersen et al. (2006) (Ped06). Original participating systems can be found in these works.

5 Discussion

Single v.s. multiple knowledge sources As shown in Table 6, considering the best performances across all feature enrichment strategies and feature sets, the proposed method successfully harnessed different knowledge sources and improved over the baselines using single knowledge sources by 0.02 ~ 0.11. The biggest improvement (0.11) is on a domain-specific dataset, on which the method based on single knowledge source performed poorly in terms of coverage and accuracy. The best enrichment strategy that has consistently improved the baselines is *wk-4F+wn-4F, Integration* (Table 4 v.s. Table 3). With features enriched from *multiple* knowledge sources, the method also consistently improved over their *corresponding single*-source features on all datasets, except MeSH36, on which *wk-4F+wn-4F, Combination* (Table 4) slightly reduced the accuracy obtained with *wk-4F* (Table 3) only.

The large proportion of uncovered word pairs using WordNet is due to its lack of coverage of specialized lexicons, and sparser semantic content. For example, of all 115 distinctive terms in the Ped30 and MeSH36 datasets, 30% are not included in WordNet. And of all 447 distinctive words in all general domain datasets, only 69% have multiple synonyms. Features such as *attributes* and "*see also*" are present for less than 20 words. This is the reason that some approaches using WordNet (e.g., Agirre et al., 2009) require a graph of all WordNet lexicons to be built, thus intermediate words may "bridge" input words even if they do not connect directly by their features. Nevertheless, the improvement in accuracy and 100% coverage after harnessing both knowledge sources suggests that they complement each other well. On one hand, Wikipedia brings its strength in domain and content coverage; on the other hand, WordNet brings useful semantic evidences for words that are covered.

Concept mapping and feature enrichment methods While the set overlap based method for cross-source concept mapping using the reference knowledge source concepts is simple and proved successful, the accuracy of mapping and its correlation with the accuracy of the SR method was not studied. This will be explored in the future. Also, alternative mapping methods will be investigated. For example, Toral and Muñoz (2006) describe a different method of mapping Wikipedia articles to WordNet synsets; one could also adopt a simple disambiguation process to select the best candidate concept from each knowledge source suited for the input word pairs, whereby cross-source concept mapping becomes straightforward. In terms of feature enrichment strategies, there is no strong indication (Table 6) of which (*feature combination v.s. integration*) is more effective, although the system consistently outperforms the baselines (Table 4 v.s. Table 3) with the *wk-4F+wn-4F, Integration* strategy.

Feature diversification v.s. unification Table 5 suggests that in most cases, differentiating feature types leads to better results than merging them uniformly, despite the knowledge sources used. This is consistent with the findings by Zhang et al. (2010). This can be understandable since although unifying feature types effectively increases possibility of sharing features, equally, this may also increase the proportion of noisy features. For example, consider the Wikipedia article of "Horse" (animal), which has a category label "livestock"; and the article "Famine", which has an outgoing link "livestock" (in a sentence describing diseases that caused decline of livestock production). By differentiating the feature types "*has_category*" and "*has_outlink*", the two concepts will not be connected even if they both have the same word "livestock" in their feature representation. However, using a bag-of-words representation where feature types are undistinguished, the strength of their relatedness is boosted by sharing this word, which may be uninteresting in this occasion.

Compared against state-of-the-art, the proposed method has achieved promising results. Overall, by harnessing different knowledge sources, the method achieves, and in many cases, outperforms state-of-the-art. In the general domain, it outperforms state-of-the-art on three out of four datasets. It is worth noting that all methods based on WordNet generally have poor performance on the Fin153 and Fin200 datasets (Table 7). Despite the heterogeneity in these datasets, this may also relate to the quality of the feature space generated with WordNet. In fact methods using Wikipedia perform better on these datasets. With enriched features from both knowledge sources, the accuracies are further improved.

In the biomedical domain, the proposed method outperforms state-of-the-art on one dataset and produces competitive results on others. Note that all other methods exploit domain-specific ontologies and corpora. The *Ped06 best* and *Ped06 second* methods also depend on a corpus of one million documents. These results further confirmed the benefits of our method: harnessing knowledge from general-purpose knowledge sources of limited domain coverage, it is possible to achieve results that rival methods based on well-curated and specially tailored domain-specific knowledge sources. This is an encouraging finding. Although there are abundant resources in the biomedical domain for this type of tasks, such resources may be scarce in other domains and are expensive to build. However, the results suggest that the proposed method offers a more affordable approach that provides reasonable coverage and quality, even if individual general knowledge sources may be limited in themselves.

Generality of the method The proposed method represents features extracted from different knowledge sources in a generic manner, which facilitates cross-source feature enrichment and requires generic algorithm computation. As discussed in Section 3, semantic evidence of words and concepts may be extracted from different knowledge sources in different ways, while harnessed in the generic model. In contrast, other methods using multiple knowledge sources (e.g., Han and Zhao, 2010; Tsang and Stevenson, 2010) introduce algorithms that are bound to the knowledge sources, which may limit their adaptability and portability.

6 Conclusion

This paper introduced a generic method of harnessing different knowledge sources to compute semantic relatedness. We have shown empirically that different knowledge sources contain complementary semantic evidence, which, when combined together under a uniform model, can improve the accuracy of SR methods. Moreover, we have demonstrated its robustness in dealing with knowledge sources of different quality and coverage. Several remaining issues will be studied in the future. First, additional knowledge sources will be studied, particularly unstructured corpora and domain-specific resources. The experiments have shown that although harnessing different knowledge sources achieved encouraging results on biomedical datasets, they are still far from being perfect. While it should be appreciated that the results are obtained using only general purpose knowledge sources, it would be interesting to investigate whether harnessing domain specific knowledge sources (where available) further improves the performance. Second, different methods of concept mapping will be studied. We will also design methods for assessing the quality of mapping, and analyze their correlations with the SR methods. Third, analyses will be carried out to uncover the differences between feature combination and integration that have led to different accuracies.

Acknowledgments

Part of this research has been funded under the EC 7th Framework Program, in the context of the SmartProducts project (231204).

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In Proceedings of NAACL'09

Batet, M., Sánchez, D., Valls, A. 2010. An ontology-based measure to compute semantic similarity in biomedicine. In Journal of Biomedical Informatics, **44**(1), 118-125

Chen, H., Lin, M., Wei, Y. 2006. Novel association measures using web search with double checking. Proceedings of COLING'06-ACL'06, pp. 1009-1016

Cilibrasi, R., Vitanyi, P. 2007. The Google Similarity Distance. In IEEE Transactions on Knowledge and Data Engineering. 19(3), 370-383

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. In ACM Transactions on Information Systems, 20 (1), pp. 116 – 131

Gabrilovich, E., Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In proceeding of IJCAI'07

Gouws, S., Rooyen, G., Engelbrecht, H. 2010. Measuring conceptual similarity by spreading activation over Wikipedia's hyperlink structure. Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources

Halavais , A. 2008. An Analysis of Topical Coverage of Wikipedia. Journal of Computer-Mediated Communication, **13**(2)

Han, X., Zhao, J. 2010. Structural semantic relatedness: a knowledge-based method to named entity disambiguation. In the 48th Annual Meeting of the Association for Computational Linguistics.

Harrington, B. 2010. A semantic network approach to measuring relatedness. In Proceedings of COLING' 10

Hirst, G., and St-Onge, D. 1998. Lexical chains as representation of context for the detection and correction malapropisms. In Christiane Fellbaum (ed.), WordNet: An Electronic Lexical Database and Some of Its Applications, pp. 305–332. Cambridge, MA: The MIT Press.

Holloway, T., Bozicevic, M., Börner, K. 2007. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. In Journal of Complexity, Special issue on Understanding Complex Systems, **12**(3), 30-40

Jiang, J. and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the International Conference on Research in Computational Linguistics, pp. 19-33

Leacock, C., Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), WordNet. An Electronic Lexical Database, Chp. 11, pp. 265-283.

Lin, D. 1998. An information-theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296-304

Lord, P., Stevens, R., Brass, A., Goble, C. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. In Bioinformatics, **19**(10), pp. 1275–1283

Matsuo, Y., T. Sakaki., K., Uchiyama, M., Ishizuka. 2006. Graph-based word clustering using a web search engine. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.542-550

Miller, G., Charles, W. 1991. Contextual correlates of semantic similarity. In Language and Cognitive Processes, 6(1): 1-28

Pan, F., Farrell, R. 2007. Computing semantic similarity between skill statements for approximate matching. In Proceedings of NAACL-HLT'07, pp. 572-579

Patwardhan, S., Pedersen, T. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together

Pedersen, T., Pakhomov, S., Patwardhan, S., Chute, C. 2006. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics **40**(3), 288-299

Pekar, V., Staab, S. 2002. Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. Proceedings of COLING'02. pp. 786-792

Pesquita, C., Faria, D., Bastos, H., Falcão, A., Couto, F. (2007). Evaluating GO-based Semantic Similarity Measures. ISMB/ECCB 2007 SIG Meeting Program Materials, International Society for Computational Biology 2007

Petrakis, E., Varelas, G., Hliaoutakis, A., Raftopoulou, P. 2006. Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies. In 4th Workshop on Multimedia Semantics (WMS'06), pp. 44-52.

Pirro, G. 2009. A semantic similarity metric combining features and intrinsic information content. In Data and Knowledge Engineering, **68**(11), pp. 1289-1308

Pozo A., Pazos F., Valencia, A. 2008. Defining functional distances over gene ontology. In BMC Bioinformatics **9**, pp.50

Rada, R., Mili, H., Bicknell, E., Blettner, M. 1989. Development and application of a metric on semantic nets. In IEEE Transactions on Systems, Man and Cybernetics **19**(1), pp.17-30

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of IJCAI-95, pp. 448-453

Riensche, R., Baddeley, B., Sanfilippo, A., Posse, C., Gopalan, B. 2007. XOA: Web-Enabled Cross-Ontological Analytics. IEEE Congress on Services, pp. 99-105

Rowe, M., Ciravegna, F. 2010. Disambiguating identity web references using Web 2.0 data and semantics. M Rowe and F Ciravegna. The Journal of Web Semantics.

Rubenstein, H., Goodenough, J. 1965. Contextual correlates of synonymy. In Communications of the ACM, 8(10):627-633

Seco, N., and Hayes, T. 2004. An intrinsic information content metric for semantic similarity in WordNet. In Proceedings of the 16th European conference on Artificial Intelligence

Strube, M., Ponzetto, S. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of the 21st national conference on Artificial intelligence (AAAI)

Toral, A., Muñoz, R. 2006. A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by using Wikipedia. In Proceedings of Workshop on New Text, ACL'06.

Tsang, V., Stevenson, S. 2010. A graph-theoretic framework for semantic distance. In Journal of Computational Linguistics, 36(1).

Wojtinnek, P., Pulman, S. 2011. Semantic relatedness from automatically generated semantic networks. In Proceedings of the Ninth International Conference on Computational Semantics (IWCS'11)

Wu, Z. Palmer, M. 1994. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133-138

Wu, H., Su, Z., Mao, F., Olman, V., Xu, Y. 2005. Prediction of functional modules based on comparative genome analysis and gene ontology application. Nucleic Acids Research, **33**, pp. 2822–2837.

Yazdani, M., Popescu-Belis, A. 2010. A random walk framework to compute textual semantic similarity: a unified model for three benchmark tasks. IEEE Fourth International Conference on Semantic Computing (ICSC), pp. 424-429

Ye, P., Peyser, B., Pan, X., Boek, J., Spencer, F., Bader, J. 2005. Gene function prediction from congruent synthetic lethal interactions in yeast. In Molecular system biology

Yeh, E., Ramage, D., Manning, C., Agirre, E., Soroa, A. 2009. WikiWalk: random walks on Wikipedia for semantic relatedness. In Proceedings of the TextGraphs-4, Workshop on Graph-based Methods for Natural Language Processing, ACL2009

Zesch, T., and Gurevych, I. 2007. Analysis of the Wikipedia category graph for NLP applications. In Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007), pp. 1–8

Zesch, T., Gurevych, I. 2010. Wisdom of crowds versus wisdom of linguists: measuring the semantic relatedness of words. In Journal of Natural Language Engineering, **16**, pp. 25-59

Zhang, Z., Gentile, A., Xia, L., Iria, J., Chapman, S. 2010. A random graph walk based approach to compute semantic relatedness using knowledge from Wikipedia. In Proceedings of LREC'10.