An Ordered Relatedness Metric for Relevance Identification

Lakshmi Ramachandran, Edward F. Gehringer
North Carolina State University University
{lramach,efg}@ncsu.edu

Abstract—In this paper we introduce a WordNet relations-based metric to determine semantic relatedness. Semantic relatedness is used to identify the degree of relevance between a review’s text and a submission’s text in order to determine whether the review pertains to the right submission. We use only WordNet since using additional corpuses or knowledge resources to determine similarity would be time consuming, especially when the metric is used to perform token-based pairwise comparison across documents. We compare our semantic relatedness metric with path and content-based measures that use only WordNet. We show that our metric is better than the other relatedness metrics at identifying relevance of academic reviews from Expertiza, a collaborative web-based learning application. We also show that our semantic relatedness metric performs better than the other metrics on the WordSim353 and Rubenstein & Goodenough datasets.

INTRODUCTION

Several applications in natural language processing and related areas involve comparison of similar-meaningd texts. An approach that relies only on a simple lexical match across compared texts may not be effective at classifying or clustering semantically related texts. In this paper we use a meaning-based similarity metric to determine relatedness across two tokens, and consequently across two documents.

Current approaches to identifying semantic relatedness use knowledge resources such as Wikipedia [1]. When comparing two terms, Wikipedia articles or concepts containing the terms to be compared are queried from over 3 million articles (≈ 3GB in size) [2]. Current techniques are also burdened by time-consuming preprocessing techniques [3]. Although Wikipedia has been shown to perform well at identifying semantic relatedness, its large size and querying time makes it difficult for smaller applications that do not have access to large clusters of parallel computers to adopt it as a knowledge resource.

Expensive preprocessing may also have scalability issues, especially when comparing texts containing hundreds of tokens. Apart from being time-consuming, Wikipedia-based approaches seem to work well when noun entities (topics) are compared.

Consider the task of review relevance identification. Reviews are text-based feedback provided by reviewers to authors. One of the factors for determining review quality is the degree of relevance between a review and the submission it is written for. We calculate relevance by identifying lexico-semantic matches between the review and submission texts [4]. Sample review and submission texts are provided in Fig. 1.

Relevance identification involves checking for paraphrases [4], i.e., lexical changes, word-order changes, use of synonyms, definitions or examples of tokens etc. Therefore, while assessing relevance, review and submission texts cannot be reduced to vectors of concepts (as done by Gabrilovich [6]). A comparison between reviews and submissions would involve checking for relatedness across verb, adjective or adverbial-forms, checking for cases of nominalizations (noun form of adjectives) etc. In this paper we introduce a unique WordNet relations-based matching approach to determine relevance, as an improvement over conventionally used path [7] and content-based measures [8].

WordNet is a widely used resource for measuring similarity. WordNet is a network of nouns, verbs, adjectives and adverbs, which are grouped into synsets (synonymous words), and linked by lexical relations [9]. WordNet does not perform as well as Wikipedia due to its limitations in terms of the domains it covers, and its lack of real-world knowledge. However WordNet is faster to query and involves no additional preprocessing. WordNet also allows comparison across different word forms.

We demonstrate that our metric outperforms conventional WordNet-based approaches while identifying semantic relatedness across terms on a review-submission dataset, as well as on commonly used datasets such as WordSim353 (WS353) [10] and Rubenstein & Goodenough (RG65) [11].
EXISTING MODELS

Semantic relatedness measures can be classified as follows: (1) Information content or path-based measures and (2) Wikipedia (concepts)-based relatedness measure. This section discusses some existing work in each of these categories.

Information content and path-based measures

Information content of a word is calculated in terms of its probability of occurrence in a corpus. Path-based measures use the distance between two nodes \( s_1 \) and \( s_2 \) in a taxonomy that are representative of the words \( w_1 \) and \( w_2 \) respectively, to determine relatedness. Jiang and Conrath use information content (IC) to determine relatedness between terms [8].

Leacock and Chodorow [12] determine relatedness in terms of the shortest distance (length) between two nodes in a taxonomy. Wu and Palmer [7] determine relatedness between tokens in terms of the depth of their least common subsumer and the depth of their nodes (\( s_1 \) and \( s_2 \)) in the taxonomy.

Wikipedia-based measures

Strube and Ponzetto compare the usefulness of both WordNet (path and information-content based measures) and Wikipedia in identifying semantic similarity across tokens or phrases [1]. Strube and Ponzetto found that WordNet did not suffer from limited coverage, since only 2 out of the 353 words in the WS353 dataset were absent in WordNet, as opposed to 13 absent words in the case of Wikipedia.

Gabrilovich et al. represent text as vectors of concepts extracted from Wikipedia (which are predominantly nouns and named entities) [6]. This type of approach may be suited for tasks such as classifying or identifying topics among texts that are abounding in named entities (e.g. news articles). Their approach involves the extraction and processing of 3 million URLs and 70GB of data.

Agirre et al.’s approach to identify relatedness involves generation of a personalized page-rank graph for every pair of compared words [3]. Generating a graph for every pair of words may not be feasible when identifying similarities between two large documents.

RELATIONS-BASED ORDERING TO DETERMINE SIMILARITY

Similar-meaninged texts may be paraphrases of each other i.e., the words and phrases in one text could be mapped to those in another text by means of a semantic relatedness metric. Researchers have found the use of synonyms, nominalizations (changing adjective to nouns e.g. using goodness instead of good), explanations of the meaning of tokens (e.g. definitions or examples) etc. to be common among paraphrases [5]. Some of our semantic relations are motivated by research in paraphrase recognition.

1) \( v \) and \( w \) are \textbf{exactly} the same.

2) \( v \) and \( w \) are \textbf{synonymous}, or if \( v \) is a nominalization of \( w \) or vice versa.

3) \( v \) is a \textbf{hyponym} of \( w \) (i.e., \( v \) is more generic than the token \( w \)) or vice versa. Or \( v \) is a \textbf{hypernym} of \( w \) (i.e., \( v \) is a more specific form of \( w \)) or vice versa.

4) \( v \) is a \textbf{meronym} of \( w \) (i.e., \( v \) is a sub-part of \( w \)) or vice versa. Or \( v \) is a \textbf{holonym} of \( w \) (i.e., \( v \) contains \( w \) as a sub-part) or vice versa. For example, “arm” is a meronym of the token “body”, and “body” is the holonym of the term “arm”.

Apart from the paraphrase-based relations, we also use a distance-based relation to determine relatedness between the synsets of two compared tokens. We also use context-based matches to help identify whether two texts’ tokens are used in similar contexts, i.e., accompanied by similar words. This may provide us with some information on the proximity of the two tokens in terms of its neighboring words. resume

1) If \( v \) and \( w \) have \textbf{common parents} (excluding generic parents such as “object”, “entity”, “organism” etc.), the normalized distance gives the similarity between the two tokens. Distance is scaled in the range of 0 to 6 (the lowest and highest values of our semantic metrics). If the scaled value is \( > 0 \), then we say that a match exists between \( v \) and \( w \).

2) An effective word-sense disambiguation technique was suggested by Lesk [13]. Lesk identifies matches between the definition of a word, whose sense is to be determined and the word, whose sense is already known. We use \textbf{overlapping definitions} as a metric to determine possible context-based similarity across tokens. For instance, tokens “quantity” and “enough” have overlapping definitions—the word “adequate” in quantity’s definition—an adequate or large amount, and enough’s definition—an adequate quantity. We also find \textbf{overlaps across examples} of the words.

The tokens are assigned a distinct or non-match if none of the above relatedness matches apply to the pair of compared words. resume

1) \( v \) and \( w \) contain \textbf{distinct} tokens or phrases.

The listed set of matches help capture relatedness between tokens based on paraphrase-based (synonyms etc.), distance-based (comm or context-based matches. A combination of the \textit{token} and its 	extit{part of speech} (POS) information is used to identify similarity between tokens. Relatedness between two terms \( v \) and \( w \)--\textit{match}(\( v \), \( w \)) is one of those listed below. Each of these types of matches is given a weight value depending on the importance of the match. Similarity matches are assigned values in the range of 0–6, since there are 7 different types of matches ranging from distinct or non-match (weight value of 0) to best or exact match (weight value of 6).

RELEVANCE IDENTIFICATION TASK

Review relevance is measured by comparing a review’s text with a submission’s text. Review and submission texts are represented using graphs. Ramachandran and Gehringer [14] have described steps involved in generating word order graphs. Graph vertices represent noun, verb, adjective or adverbial words or phrases in a text, and edges represent relationships between vertices.
Relevance is measured as the average of the best matches for the vertex and edge structures of the review and submission graphs. Vertices and edges capture context and structure information (Fig. 2(a)). Edge structures capture ordering information, which helps us identify whether the reviews contain paraphrases or structural changes in the text (Fig. 2(b)). We identify structural changes such as active to passive voice change by comparing edges in same and different orders. When we compare the edge “paper–presented” with edge “presentation–included”, we compare “paper” with “included” and “presented” with “presentation”. Token “presentation” is the nominalization of “presented”, as a result of which a match is identified between the two edges. A match between edges indicates a stronger relationship between the review and submission texts.

Relevance of a review to a submission is calculated using the formula in Equation 1, where \( \text{relevance}(S, R) \) is the similarity between submission and review graphs \( S \) and \( R \). The aim is to identify the best semantic match for every vertex or edge in the review with a vertex or edge in the submission.

\[
\text{relevance}(S, R) = \frac{1}{2|V_S|} \sum_{v_r \in V_R} \arg\max_{v_s \in V_S} \{ \text{match}_v(v_s, v_r) \} + \frac{1}{2|E_S|} \sum_{e_r \in E_R} \arg\max_{e_s \in E_S} \{ \text{match}_e(e_s, e_r) \}
\]

\( V_S, V_R \) are the vertices, and \( E_S \) and \( E_R \) are the edges in the submission and review graphs respectively. The value of \( \text{match}_v \) is identified using our semantic relatedness metric. For instance, if two tokens have an exact match the \( \text{match}_v \) value is 6, or if they have a synonym match the value is 5. Edge match \( \text{match}_e \) is the average of the matches of the edge’s vertices.

**Experiments**

**Evaluating relevance**

In this section we compare the performance of the different types of semantic relatedness metrics in determining review relevance. To ensure that our relations-based metric is a suitable similarity identification metric, we compare our approach’s results with other relatedness measures that use only WordNet (and no external knowledge resources or corpuses). We use WordNet to implement the information content-based measure proposed by Jiang and path-based measures proposed by Wu and Leacock. Our aim with this evaluation is to show the usefulness of our relations-based metric in identifying relevance. We use precision, recall and \( f \)-measure to evaluate the metrics’ performance.

We perform our experiments on review and submission data from two assignments completed using Expertiza [15]. Expertiza is a collaborative learning application, which includes peer-reviewing functionality. In a review-submission comparison reviews are compared with their respective submission texts. In addition, reviews are compared with other submission texts to include some explicit non-relevant data points. We use 292 data points containing an equal distribution of relevant and non-relevant review-submission pairs for evaluation. Around 50% of the review-submission data was randomly selected and annotated by two annotators. The ratings had a 78.3% agreement and a Cohen’s Kappa value of 0.6.

The results of using different relatedness metrics in determining relevance are listed in Table I. Our semantic relatedness metric produces the highest \( f \)-measure.

**Product reviews from Amazon:** In order to study the generalizability of our semantic relatedness metric on other datasets, we study its performance in determining relevance of product reviews from the Amazon dataset [16]. Relevance may be applied to study how relevant product reviews are to a description of the product.

The Amazon reviews dataset contains tags of product features that are described (if any) in the reviews. We use these tags to determine whether a review is relevant or not to a product’s description. Reviews that discuss product features are tagged as relevant, and those without feature tags are marked as non-relevant. We randomly selected 400 product reviews containing an equal number of relevant and non-relevant review-submission pairs for evaluation. Around 50% of the review-submission data was randomly selected and annotated by two annotators. The ratings had a 78.3% agreement and a Cohen’s Kappa value of 0.6.

The results of using different relatedness metrics in determining relevance are listed in Table I. Our semantic relatedness metric produces the highest \( f \)-measure.
We have shown that the relative order, established across the
tokens, helps capture the similarity between
tokens. We used exponential values (1, 2, 4, 8, 16, 32, 64) and random values (2, 10, 23, 40, 47, 50, 52) as weights for the WordNet metrics.

Results listed in Table II indicate that the correlations achieved by the exponential and random weight values are comparable to those achieved by the [0–6] weight scale. Irrespective of the values, the relative ordering of the WordNet metrics used by our approach produce high correlations. Therefore the values of the weights are not as important as their relative order.

Correlations 0.43 and 0.47 achieved by our metric on the WS353-full and test datasets indicate that our metric has a positive correlation with human annotations. These correlations are greater than those achieved by path and information content based measures [1].

Our approach produces a correlation of 0.83 and Leacock’s metric produces a correlation of 0.86 on the RG65 dataset. Leacock’s metric performs better than ours for RG65 since their metric identifies the distance between two synsets, while our approach looks for specific types of relations. As a result even when two tokens such as “food” and “rooster” have synsets that are distantly related they get a similarity value (of 7.19 for this example) using Leacock’s metric. However, if none of the specific types of matches apply to the compared tokens, then they get a distinct or no-match value of 0 using our approach. This offsets our approach’s correlations with human-provided values.

Correlations produced by WordNet-based measures on the RG65 dataset are greater than the correlation of 0.82 achieved by a Wikipedia-based approach proposed by Gabrilovich et al. [6].

**Conclusion**

We introduce a unique relations-based semantic relatedness measure. We use features such as exact matches, synonymy, hypernymy, hyponymy, holonymy, meronymy, common parents and overlapping definitions to determine a review’s relevance to a submission. We demonstrate that our approach is suited for identifying a review’s relevance to a submission. We have shown that the relative order, established across the different semantic relations, outperforms path and content-based measures that use only WordNet. To improve the process of relevance identification in the future, we plan on studying the impact of using ontologies in the educational domain, which may not involve excessive preprocessing or data transformation.

**Acknowledgment**

This work has been supported by the NSF grant # 0942279.

**References**


**Table II: Evaluating the relative ordering of the WordNet match types**

<table>
<thead>
<tr>
<th>WordNet metric</th>
<th>WS353-Full</th>
<th>WS353-Test</th>
<th>RG65</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–6 scale</td>
<td>0.43</td>
<td>0.47</td>
<td>0.83</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.41</td>
<td>0.45</td>
<td>0.83</td>
</tr>
<tr>
<td>Random values</td>
<td>0.42</td>
<td>0.46</td>
<td>0.83</td>
</tr>
<tr>
<td>Wu [1]</td>
<td>0.30</td>
<td>0.28</td>
<td>0.82</td>
</tr>
<tr>
<td>Leacock [1]</td>
<td>0.34</td>
<td>0.35</td>
<td>0.86</td>
</tr>
<tr>
<td>Jiang</td>
<td>0.31</td>
<td>0.36</td>
<td>0.54</td>
</tr>
</tbody>
</table>