

ClaC: Semantic Relatedness of Words and Phrases

Reda Siblini

Concordia University
1400 de Maisonneuve Blvd. West
Montreal, Quebec, Canada, H3G 1M8
r_sibl@encs.concordia.ca

Leila Kosseim

Concordia University
1400 de Maisonneuve Blvd. West
Montreal, Quebec, Canada, H3G 1M8
kosseim@encs.concordia.ca

Abstract

The measurement of phrasal semantic relatedness is an important metric for many natural language processing applications. In this paper, we present three approaches for measuring phrasal semantics, one based on a semantic network model, another on a distributional similarity model, and a hybrid between the two. Our hybrid approach achieved an F-measure of 77.4% on the task of evaluating the semantic similarity of words and compositional phrases.

1 Introduction

Phrasal semantic relatedness is a measurement of how multiword expressions are related in meaning. Many natural language processing applications such as textual entailment, question answering, or information retrieval require a robust measurement of phrasal semantic relatedness. Current approaches to address this problem can be categorized into three main categories: those that rely on a knowledge base and its structure, those that use the distributional hypothesis on a large corpus, and hybrid approaches. In this paper, we propose supervised approaches for comparing phrasal semantics that are based on a semantic network model, a distributional similarity model, and a hybrid between the two. Those approaches have been evaluated on the task of semantic similarity of words and compositional phrases and on the task of evaluating the compositionality of phrases in context.

2 Semantic Similarity of Words and Compositional Phrases

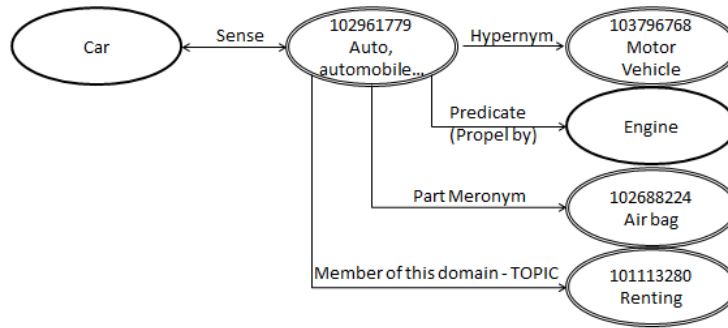
The semantic similarity of words and compositional phrases is the task of evaluating the similarity of a word and a short phrase of two or more words; for example, the word *Interview* and the phrase *Formal Meeting*. In the next section we present our semantic network model for computing phrasal semantic relatedness between a word and a phrase, followed by a distributional similarity model, that we evaluate on the task of semantic similarity of words and compositional phrases.

2.1 Semantic Network Model

Knowledge-based approaches to semantic relatedness use the features of the knowledge base to measure the relatedness. One of most frequently used semantic network is the Princeton's WordNet (Fellbaum, 1998) which groups words into synonyms sets (called synsets) and includes 26 semantic relations between those synsets, including: hypernymy, hyponymy, meronymy, entailment . . .

To measure relatedness, most of those approaches rely on the structure of the semantic network, such as the semantic link path, depth (Leacock and Chodorow, 1998; Wu and Palmer, 1994), direction (Hirst and St-Onge, 1998), or type (Tsatsaronis et al., 2010). Our phrasal semantic relatedness approach is inspired from those methods. However, our approach is based on the idea that the combination of the least costly types of relations that relate one concept to a set of concepts are a suitable indicator of their semantic relatedness. The type of relations considered includes not only the hy-

Figure 1: Example of the semantic network around the word *car*.



ponym/hypernym relations but also all 26 available semantic relations found in WordNet in addition to relations extracted from each of the eXtended WordNet (Harabagiu et al., 1999) synset’s logical form.

To implement our idea, we created a weighted and directed semantic network based on the relations of WordNet and eXtended WordNet. We used WordNet’s words and synsets as the nodes of the network. Each word is connected by an edge to its synsets, and each synset is in turn connected to other synsets based on the semantic relations included in WordNet. In addition each synset is connected by a labeled edge to the predicate arguments that are extracted from the eXtended WordNet synset’s logical form. Every synset in the eXtended WordNet is related to a logical form, which contains a set of predicate relations that relates the synset to set of words. Each predicate in this representation is added as an edge to the graph connecting the synset to a word. For example, Figure 1 shows part of the semantic network created around the word *car*. In this graph, single-line ovals represent words, while double-line ovals represent synsets.

To compute the semantic relatedness between nodes in the semantic network, it is necessary to take into consideration the semantic relation involved between two nodes. Indeed, WordNet’s 26 semantic relations are not equally distributed nor do they contribute equally to the semantic relatedness between concept. In order to indicate the contribution of each relation, we have classified them into seven categories: *Similar*, *Hypernym*, *Sense*, *Predicate*, *Part*, *Instance*, and *Other*. By classifying WordNet’s relations into these classes, we are able to weight

the contribution of a relation based on the class it belongs to, as opposed to assigning a contributory weight to each relations. The weights were assigned by manually comparing the semantic features of a set of concepts that are related by a specific semantic relations. Table 1 shows the seven semantic categories that we defined, their corresponding weight, and the relations they include. For example the category *Similar* includes WordNet’s relations of *entailment*, *cause*, *verb group*, *similar to*, *participle of verb*, *antonym*, and *pertainym*. This class of relations has the most common semantic features when comparing two concepts related with any of those relations and hence was assigned the lowest weight¹ of 1. All the 26 relations in the table are the ones found in WordNet, for the exception of the predicate (and inverse predicate) relations which are the predicate relations extracted from the eXtended WordNet. This can be seen in Figure 1, for example, where the word *car* is related to the word *Engine* with the *Predicate* relation extracted from the eXtended WordNet logical form and more specifically the predicate *propel by*.

The computation of semantic relatedness between a word and a compositional phrase is then the combination of weights of the shortest weighted path² in the weighted semantic network between that word and every word in that phrase, normalized by the maximum path cost.

¹The weight can be seen as the cost of traversing an edge; hence a lower weight is assigned to a highly contributory relation.

²The shortest path is based on an implementation of Dijkstra’s graph search algorithm (Dijkstra, 1959)

Category	Weight	Semantic Relations in WordNet or xWordnet
<i>Similar</i>	1	similar to, pertainym, participle of verb, entailment, cause, antonym, verb group
<i>Hypernym</i>	2	hypernym, instance hypernym, derivationally related
<i>Sense</i>	4	lemma-synset
<i>Predicate</i>	6	predicate (extracted from Extended WordNet)
<i>Part</i>	8	holonym (instance, member, substance), meronym (instance, member, substance), inverse predicate (extracted from Extended WordNet)
<i>Instance</i>	10	hyponym, instance hyponym
<i>Other</i>	12	attribute, also see, domain of synset (topic, region, usage), member of this domain (topic, region, usage)

Table 1: Relations Categories and Corresponding Weights.

Figure 2 shows an extract of the network involving the words *Interview* and the phrase *Formal Meeting*. For the shortest path from *Interview* to *Formal*, the word *Interview* is connected with a *Sense* relation to the synset #107210735 [*Interview*]. As indicated in Table 1, the weight of this relation is defined as 4. This synset is connected to the synset *Examination* through a *Hypernym* relation type with a weight of 2, which is connected to the word *Formal* with a predicate (IS) relation of weight 6. Overall, the sum of the shortest path from *Interview* to *Formal Meeting* is hence equal to the sum of the edges shown in Figure 1 ($4+2+6+4+6+4+6 = 32$). By normalizing the sum to the maximum, In our approach, 24 is maximum path cost after which we assume that two words are not related (which we assume to be traversing two times maximum weighted path, $2 * \text{maximum path weight of } 12$) and 8 is the minimum number of edges between 2 words (which is equal to traversing from the word to itself, $2 * \text{sense weight of } 4$). Taking into consideration the number of words in the phrase, the semantic relatedness will be $(24*2 - (32-8*2))/24*2 = 66.7\%$. In the next section, we will introduce our distributional similarity model.

2.2 Distributional Similarity Model

Distributional similarity models rely on the distributional hypothesis (Harris, 1954) to represent a word by its context in order to compare word semantics. There are various approach for the selection, repre-

sentation, and comparison of contextual data. Most use the vector space model to represent the context as dimensions in a vector space, where the feature are frequency of co-occurrence of the context words, and the comparison is usually the cosine similarity. To go beyond lexical semantics and to represent phrases, a compositional model is created, some use the addition or multiplication of vectors such as Mitchell and Lapata (2008), or the use of tensor product to account for word order as in the work of Widdows (2008), or a more complex model as the work of Grefenstette and Sadrzadeh (2011). In our model, we are inspired by those various work, and more specifically by the work of Mitchell and Lapata (2008). The compositional model is based on phrase words vectors addition, where each vector is composed of the collocation pointwise mutual information of the word up to a window of 3 words left and right of the main word. The corpus used to collect the features and their frequencies is the Web 1TB corpus (Brants and Franz, 2006). For the *Interview* to *Formal Meeting* example, the vector of the word *interview* is first created from the corpus of the top 1000 words collocating *interview* between the window of 1 to 3 words with their frequencies. A similar vector is created for the word *Formal* and the word *Meeting*, the vector representing *Formal Meeting* is then the addition of vector *Formal* to vector *Meeting*. The comparison of vector *Interview* to vector *Formal Meeting* is then the cosine of both vectors.

2.3 Evaluation

We evaluated our approaches for word-phrase semantic relatedness on the SemEval task of evaluating phrasal semantics, and more specifically on the sub-task of evaluating the semantic similarity between words and phrases. The task provided an English dataset of 15,628 word-phrases, 60% annotated for training and 40% for testing, with the goal of classifying each word-phrase as either positive or negative. To transform the semantic relatedness measure to a semantic similarity classification one, we first calculated the semantic relatedness of each word-phrase in the training set, and used JRip, WEKA's (Witten et al., 1999) implementation of Cohen's RIPPER rule learning algorithm (Cohen and Singer, 1999), in order to learn a set of rules that can differentiate between a positive semantic similarity and a negative one. The classifier resulted in rules for the semantic network model based relatedness that could be summarized as follows: *If the semantic relatedness of the word-phrase is over 61% then the similarity is positive, otherwise it is negative.* So for the example *Interview - Formal meeting*, which resulted in a semantic relatedness of 66.7% in the semantic network approach, it will be classified positively by the generated rule. This method was our first submitted test run to this task, which resulted in a recall of 63.79%, a precision of 91.01%, and an F-measure of 75.00% on the testing set.

For the second run, we trained the distributional similarity model using the same classifier. This resulted with the following rule that could be summarized as follows: *If the semantic relatedness of the word-phrase is over 40% then the similarity is positive, otherwise it is negative.* It was obvious from the training set that the semantic network model was more accurate than the distributional similarity model, but the distributional model had more coverage. So for our second submitted test run, we used the semantic network approach as the main result, but used the distributional model as a backup approach if one of the words in the phrase was not available in WordNet, thus combining the precision and coverage of both approaches. This method resulted in a recall of 69.48%, a precision of 86.70%, and an F-measure of 77.14% on the testing set.

For the last run, we used the same classifier

but this time we training it using two features: the semantic network model relatedness measure (SN), and the distributional similarity model (DS). This training resulted in a set of rules that could be summarized as follows: *if SN > 61% then the similarity is positive, else if DS > 40% then the similarity is also positive, and lastly if SN > 53% and DS > 31% then also in this case the similarity is positive, otherwise the similarity is negative.* This was our third submitted test run, which resulted a recall of 70.66%, a precision of 85.55%, and an F-measure of 77.39% on the testing set.

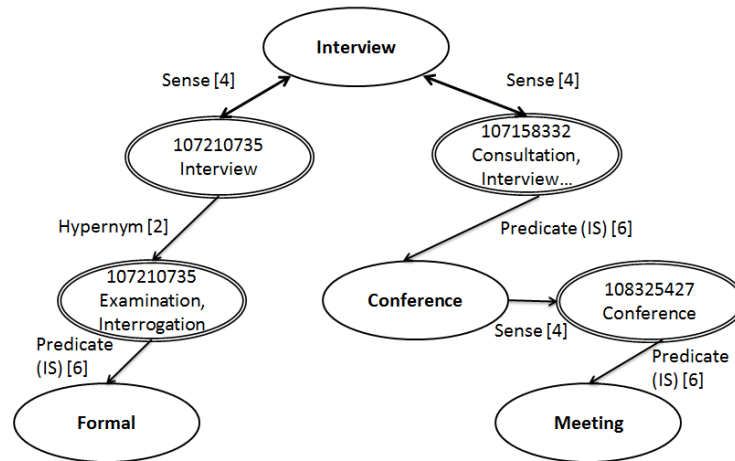
3 Semantic Compositionality in Context

The semantic compositional in context is the task of evaluating if a phrase is used literally or figuratively in context. For example, the phrase *big picture* is used literally in the sentence *Click here for a bigger picture* and figuratively in *To solve this problem, you have to look at the bigger picture.*

Our approach for this task is a supervised approached based on two main components: first, the availability of the phrases most frequent collocating expressions in a large corpus, and more specifically the top 1000 phrases by frequency in Web 1TB corpus (Brants and Franz, 2006). For example, for the phrase *big picture*, we collect the top 1000 phrases that come before and after the phrase in a corpus, those includes *look at the, see the, understand the* If the context contain any of those phrase, then this component returns 1, indicating that the phrase is most probably used figuratively. The intuition is that, the use of phrases figuratively is more frequent than their use in a literal meaning, and hence the most frequent use will be collocated with phrases that indicate this use.

The second component, is the phrase compositionality. We calculate the semantic relatedness using the semantic network model relatedness measure, that was explained in Section 2.1, between the phrase and the first content word before it and after it. The intuition here is that the semantic relatedness of the figurative use of the phrase to its context should be different than the relatedness to its literal use. So for the example, the phrase *old school* in the context *he is one of the old school* versus *the hall of*

Figure 2: Shortest Path Between the Word *Interview* and the Phrase *Formal Meeting*.



the old school, we can notice that *hall* will be more related to *old school* than the word *one*. This component will result in two features: the relatedness to the word before the phrase (SRB) and the relatedness to word after the phrase in context (SRA).

To combine both components, we evaluated our approaches on the data set presented by the SemEval task of evaluating phrasal semantics, and more specifically on the sub task of evaluating semantic compositionality in context. The data set contains a total of 1114 training instances, and 518 test instances. We use the training data and computed the three features (Frequent Collocation (FC), Semantic Relatedness word Before (SRB), and Semantic Relatedness word After (SRA), and used JRip, WEKA's (Witten et al., 1999) implementation of Cohen's RIPPER rule learning algorithm (Cohen and Singer, 1999) to learn a set of rule that differentiate between a figurative and literal phrase use. This method resulted in a set of rules that can be summarized as follows: *if FC is equal to 0 and SRB < 75% then it is used literally in this context, else if FC is equal to 0 and SRA < 75% then it is also used literally, otherwise it is used figuratively*. This method resulted in an accuracy of 55.01% on the testing set.

4 Conclusion

In this paper we have presented state of the art word-phrase semantic relatedness approaches that are based on a semantic network model, a distributional model, and a combination of the two. The

novelty of the semantic network model approach is the use of the sum of the shortest path between a word and a phrase from a weighted semantic network to calculate word-phrase semantic relatedness. We evaluated the approach on the SemEval task of evaluating phrasal semantics, once in a supervised standalone configuration, another with a backup distributional similarity model, and last in a hybrid configuration with the distributional model. The hybrid model achieved the highest f-measure in those three configuration of 77.4% on the task of evaluating the semantic similarity of words and compositional phrases. We also evaluated this approach on the subtask of evaluating the semantic compositionality in context with less success, and an accuracy of 55.01%.

Acknowledgments

We would like to thank the reviewers for their suggestions and valuable comments.

References

- Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1.
- William W Cohen and Yoram Singer. 1999. A simple, fast, and effective rule learner. In *Proceedings of the National Conference on Artificial Intelligence*, pages 335–342. John Wiley & Sons Ltd.
- Edsger W Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.
- Sanda Harabagiu, George Miller, and Dan Moldovan. 1999. Wordnet 2- a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX*, volume 99, pages 1–8.
- Zellig S Harris. 1954. Distributional structure. *Word*.
- Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet An electronic lexical database*, pages 305–332, April.
- Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.
- Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244.
- George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. 2010. Text relatedness based on a word thesaurus. *Journal of Artificial Intelligence Research*, 37(1):1–40.
- Dominic Widdows. 2008. Semantic vector products: Some initial investigations. In *To appear in Second AAAI Symposium on Quantum Interaction*, volume 26, page 28th. Citeseer.
- Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. Weka: Practical machine learning tools and techniques with java implementations.
- Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, New Mexico, June.