Combining Local and Related Context for Word Sense Disambiguation on Specific Domains

Franco Rojas-Lopez, Ivan Lopez-Arevalo and Victor Sosa-Sosa
Cinvestav-Tamaulipas, Scientific and Technological Park, Victoria, Mexico

Keywords: Semantic Similarity, Local and Related Context, Word Sense Disambiguation.

Abstract: In this paper an approach for word sense disambiguation in documents is presented. For Word Sense Disambiguation (WSD), the local and related context for an ambiguous word is extracted, such context is used for retrieve second order vectors from WordNet. Thus two graphs are built at the same time and evaluated individually, finally both results are combined to automatically assign the correct sense for the ambiguous word. The proposed approach was tested on the task #17 of the SemEval 2010 international competition producing promising results compared to other approaches.

1 INTRODUCTION

For structural and cognitive reasons, the natural language is inherently ambiguous; for example, a single lexical unit can have different meanings, this phenomenon is called polysemy[1]. When a word is polysemous one needs an algorithm to select the most appropriate meaning[2] for the given word in relation to the given context. The problem of assigning concepts to words in texts is known as Word Sense Disambiguation (WSD). A word is ambiguous when its meaning varies depending on the context in which it occurs. To date, supervised systems have obtained the best performing on WSD task, such approaches rely on the availability of sense-labeled data from which the relevant sense distinctions are learned. Unfortunately the manual creation of knowledge resource is an expensive and time consuming effort (Gale et al., 1992), furthermore labeled data are often highly domain dependent. An unsupervised approach typically refers to disambiguating word senses without the use of sense-tagged corpora.

In this paper, a knowledge-based approach on specific domain is presented, which uses unlabeled data in the disambiguation process. The approach integrates semantic information derived from an untagged corpus of a specific domain, named *related context*, and the actual occurrence context named *local context* for each ambiguous word. *Second order vectors* (Patwardhan and Ted, 2006) are extracted from WordNet and modeled as a graph of semantic relationships between synsets to perform Word Sense Disambiguation for each ambiguous word in the documents.

[1]The association of one word with two or more distinct meanings.

[2]Hereafter the terms meaning, sense, concept, and synset are used interchangeably.

2 RELATED WORK

The field of WSD is a well studied research area (Navigli, 2009) mainly because WSD is essential to have success in several linguistic tasks. Schütze and Pedersen (Schütze and Pedersen, 1995) demonstrated that WSD can be used to improve the performance of Information Retrieval tasks, enhancing the precision about 4.3%. Recently, it has been reported that graph-based methods have shown to outperform other systems (Agirre et al., 2010). Reddy (Reddy et al., 2010) used an untagged corpus from a target domain to construct a distributional thesaurus of related words; after, each ambiguous word was disambiguated using the Personalized PageRank algorithm (Agirre and Soroa, 2009). Navigli and Lapata (Navigli and Lapata, 2007) explored several measures for analy-zing the connectivity of the semantic graph structure in local and global level. They concluded that local measures perform better than global measures. The present work has a different focus based mainly on the concept of *second order vectors*, which combines semantic similarity and local context information, as

Rojas-Lopez F., Lopez-Arevalo I. and Sosa-Sosa V..
Combining Local and Related Context for Word Sense Disambiguation on Specific Domains.
DOI: 10.5220/0004046001350140
In *Proceedings of the International Conference on Data Technologies and Applications* (DATA-2012), pages 135-140
ISBN: 978-989-8565-18-1

is explained in the Section 6. The preliminary experiments show a promising response on domain-based WSD.

3 APPROACH

The proposed approach relies on the integration of two information sources, related terms together with the local context for disambiguate each ambiguous word.

3.1 Semantic Similarity

This section describe two semantic similarity measures to retrieve related terms for each ambiguous word, (mutual information and distributional hypothesis) which consist in the following three steps: (1) keyword term extraction, (2) web query, and (3) semantic similarity measures.

3.1.1 Web Query

To increase the size of the corpus provided for the SemEval, TF-IDF (Navigli et al., 2011) has been applied to retrieve keywords from untagged corpus. The first 20 keywords from the corpus were selected and combined in pairs to generate web queries of length two according to Iosif and Potamianos (Iosif and Potamianos, 2009). For example, for "*sumatra and permafrost*", the web queries were sent to several search engines to retrieve related web documents.

The increased corpus is used to retrieve related terms of each ambiguous word, which is named *related context*.

1. **Mutual Information (MI).**
   In this case, a *bag-of-words* model is described. This model assumes that the order of words has no significance (the term *"environmental evaluation"* has the same probability as *"evaluation environmental"*). Thus after of pre-processing phase, a sorted list of relevant neighbor words for each ambiguous word are retrieved from the corpus using MI (Kenneth and Patrick, 1990) (Eq. 1). The context window[3] size to retrieve related terms was defined as $2\delta+1$, $\delta=5$. Let w be the ambiguous word and t_i a term, $MI(w,t_i)$ denotes how many times the term t_i appears with word w.

$$MI(w,t_i) = log_2 \frac{f(w,t_i)}{f(w)f(t_i)} \quad (1)$$

[3]The parts that immediately precede and follow a word and clarify its meaning.

2. **Distributional Hypothesis (DH).**
   The word-context matrices are the most suited for measuring the semantic similarity of word pairs and patterns (Turney and Pantel, 2010). Thus, in addition to MI a matrix-based representation to retrieve related semantic terms with ambiguous word was tested, which is described as follows. The first step consists in retrieve multiple contexts in which an ambiguous word occurs, thereby, a parser is used to extract contextual relations. Formally a context relation or context is a tuple $<w,r,w'>$ where w is a headword occurring on some relation type r with another word w' in one or more sentences. Each occurrence extracted from raw text is an instance of a context, in this case the tuple (r,w') is an attribute of w. In this work all the grammatical relations defined by Catherine and Manning (Catherine and Manning, 2008) are used. Once that the contextual relations of each headword has been extracted from the corpus, a word-context matrix (M) is used as representation, each item (m_{ij}) in the matrix has associated a frequency. The frequency is used by weight functions to assign higher values to contexts that are more indicative of the meaning of a word. In this approach two weight functions were tested, the Eqs. 2 and 4 use the notation proposed by Richard (Richard, 2004) where $f(w,r,w')$ is the total instances of the context where w appears in and $n(^*,r,w')$ is the number of attributes that r,w' appears with. Other weight function implemented was Pointwise Mutual Information (PMI) (Zhao and Lin, 2004) (Eq. 3), which measures the strength association between an attribute f_i and a word w.

$$TF-IDF = \frac{f(w,r,w')}{n(^*,r,w')} \quad (2)$$

$$PMI(w,f_i) = log \frac{P(w,f_i)}{p(f_i)p(w)} \quad (3)$$

Once the weight matrix is obtained, the context of a word is represented as a feature vector. The similarity between two words (w_1, w_2) is computed using these vectors. The Eq. 4 computes the cosine between their feature vectors using the weight defined by the Eq. 2; here a superscript asterisk indicates that the variables are bound together as is defined by Richard (Richard, 2004). The similarity between two words using cosine of PMI is defined by Eq. 5 and uses the weight defined by Eq. 3.

$$Cosine(w_1,w_2) = \frac{\sum wgt(w_1,*_r,*_{w'})wgt(w_2,*_r,*_{w'})}{\sqrt{\sum wgt(w_1,^*,^*)^2 \sum wgt(w_2,^*,^*)^2}} \quad (4)$$

$$Sim_{CosPMI}(w_1, w_2) = \frac{\sum_{i=1}^{n} pmi(f_i, w_1) pmi(f_i, w_2)}{\sqrt{\sum_{i=1}^{n} pmi(f_i, w_1)^2} \sqrt{\sum_{i=1}^{n} pmi(f_i, w_2)^2}} \quad (5)$$

Thus, tested the techniques (MI and DH), two word lists sorted in descending order according to their semantic similarity respect to each ambiguous word are retrieved from an untagged corpus.

3.2 Local Context

Here the objective is to obtain the *local context* of each ambiguous word. *Local context* is the shortest step in this approach. In this step, the test data released by the last SemEval conference was used. SemEval is an international competition on semantic analysis systems where different system may evaluate their performance. Particularly the test data in the task #17 released by SemEval 2010 are tagged using the stanford POS tagger. Afterward, different window size were tested in the experiments to determine how many words before and after an ambiguous word *w* must be included in the context. The better resulting window size was $2\delta + 1$, with $\delta = 1$. Thus, the local context is formed by at least three words including the ambiguous word.

4 GRAPH CONSTRUCTION

Hereafter the term *related context* and *local context* are used interchangeably to denote related words and the actual occurrence context. In the literature some semantic similarity measures have been implemented to quantify the degree of similarity between two words using information drawn from the WordNet hierarchy (Pedersen et al., 2004). Particularly the Lin and Vector measures were taken into account in the conducted research. Once contexts are recovered, the senses for each word in the context are retrieved from WordNet and weighted by a semantic similarity score using the WordNet::Similarity[4] score between the senses of word *w* and the senses for each word in the context. These measures return a real value indicating the degree of semantic similarity between a pair of concepts.

Formally let $C_w = \{c_1, c_2, \cdots, c_n\}$ the set of words in the related context to an ambiguous word *w*. Let *senses(w)* be the set of senses of *w* and let *senses(c_n)* be the set of senses for a word in the context, a ranked list is returned in descending order of semantic similarity between *w* and c_n. The items that maximize this score are filtered according to a threshold $\theta = 0.35$, thus the senses in c_n closely related with the senses of *w* are retrieved. These items constitute the named *first order vectors*. For each ambiguous word, two graph are built at the same time. The representation of the graph is given by $G = (V, E, W)$ where V are the vertices (concepts), E are the edges (semantic relations) and W (a strong link between two concepts or vertices). Each recovered sense again is tagged with the Part-Of-Speech to recover the *second order vectors* for each word within the first sense. These semantic relations for senses constitute the connections in the graph. Once the semantic graph is built, its structure and links are analyzed applying the algorithms described in the following section.

[4]This is a Perl module that implements a variety of semantic similarity and relatedness measures, Ted Pedersen (Pedersen et al., 2004) http://wn-similarity.sourceforge.net/, visited January 15, 2012.

5 GRAPH-BASED MEASURES

Vertex-based centrality is defined in order to measure the importance of a vertex in the graph; a vertex with high centrality score is usually considered more highly influential than other vertex in the graph. In the approach, three algorithms have been implemented to determine which node is the most important.

- **Indegree** (Sinha and Mihalcea, 2007), the simplest and most popular measure is degree centrality. In an undirected graph the degree of the vertex is the number of its attached links; it is a simple but effective measure of nodal importance. A node is important in a graph as many links converge to it. Let V the set of vertices on the graph and v a vertex, this measure is defined by Eq. 6.

$$score(v) = \frac{indegree(v)}{|V| - 1} \quad (6)$$

- **Key Problem Player** (KPP) (Navigli and Lapata, 2007), consists in finding a set of nodes that is maximally connected to all other nodes. Here, a vertex (denoted by v and u, V is the set of vertices) is considered important if it is relatively close to all other vertices Eq. 7.

$$kpp(v) = \frac{\sum_{u \epsilon V : u \neq v} \frac{1}{d(u,v)}}{|V| - 1} \quad (7)$$

- **Personalized PageRank** (PPRank) (Agirre and Soroa, 2009), this is a modified version of the original PageRank (Brin and Page, 1998), which consists on use undirected graph with weight in

Table 1: Performance using only semantic similarity information.

Model used	Algorithm	Precision (%)	Recall (%)
Sim_{CosPMI}	kpp	9.82	9.58
	Indegree	22.58	22.03
	PPRank	7.62	7.43
Cosine	kpp	3.44	3.36
	Indegree	13.85	13.51
	PPRank	1.61	1.57
MI	kpp	2.71	2.64
	Indegree	12.17	11.87
	PPRank	1.9	1.85

Table 2: Performance using only context.

Algorithm	Precision (%)	Recall
kpp	21.33	20.81
Indegree	21.99	21.45
PPRank	1.61	1.57

Table 3: Integration of semantic similarity and context information.

Model used	Algorithm	Precision (%)	Recall (%)
Sim_{CosPMI}	kpp	31.89	31.11
	Indegree	41.27	40.27
	PPRank	11.58	11.3
Cosine	kpp	27.27	26.6
	Indegree	38.19	37.26
	PPRank	12.9	12.58
MI	kpp	26.68	26.03
	Indegree	37.17	36.26
	PPRank	13.34	13.01

Table 4: Overall results for the domain WSD of SemEval 2010.

Algorithm	Precision (%)	Recall (%)
Anup Kulkarni	51.2	49.5
Andrew Tran	50.6	49.3
Andrew Tran	50.4	49.1
Aitor Soroa	48.1	48.1
⋮	⋮	⋮
Hansen A. Schwartz	43.7	39.2
Our approach	41.2	40.2
Aitor Soroa	38.4	38.4
Radu Ion	35.1	35.0
Yoan Gutierrez	31.2	30.3
Random baseline	23.2	23.2

the edges. After running the algorithm, a score is associated with each vertex as shows the Eq. 8.

$$PR(v_i) = (1-\alpha) + \alpha * \sum_{v_j \varepsilon In(v_i)} \frac{w_{ji}}{\sum_{v_k \varepsilon Out(v_j)} w_n k} PR(v_j) \quad (8)$$

According to the literature, α is a factor which is usually set as 0.85, that is the value used in this evaluation.

6 SPECIFIC DOMAIN WORD SENSE DISAMBIGUATION

In this section the related and local context are integrated to disambiguate the words in a test document. For each word *w* to be disambiguated in the document, two graphs were built and evaluated independently, one with the local context and other with the related context, applying graph centrality algorithms (Section 5), thus two vectors were obtained as a result of these evaluations, $V_{rt} = \{x_1, x_2, \cdots, x_n\}$ and $V_{lc} = \{x_1, x_2, \cdots, x_m\}$. These vectors are integrated as shows Eq. 9 to produce a final sorted vector of synsets.

$$V_{final} = \frac{V_{rt}[x] * V_{lc}[x]}{V_{rt}[x] + V_{lc}[x]} \quad (9)$$

where *x* is a item in the vector, $V_{rt}[x] * V_{lc}[x]$ is defined as $\{V_{rt}[x] * V_{lc}[x] \mid x \varepsilon V_{rt} \cap V_{lc}\}$, and $V_{rt}[x] + V_{lc}[x]$ is defined as $\{ V_{rt}[x] + V_{lc}[x] \mid x \varepsilon V_{rt}, x \varepsilon V_{lc}\}$. The synset with the highest value is selected as the right sense for the ambiguous word *w*.

7 EXPERIMENTS AND RESULTS

In this section the obtained results for WSD on a specific domain are presented.

7.1 Test Data

For the experiments, the gold standard dataset released by SemEval 2010 (Agirre et al., 2010) was used. This dataset contains 1,398 instances of ambiguous words, 366 verbs, and 1032 nouns. For efficiency reasons, in the experiments the *related context* is formed by 5 semantic terms and *local context* as it was described above (Subsection 3.2).

7.2 Analysis

Tables 1, 2, and 3 show the results obtained with the algorithms used in the described approach. As can be seen in tables, the Indegree measure obtained the best results in the three scenarios: using only the semantic similarity, using only the context information, and combining both techniques. A best performance was obtained in the third scenario. This fact motivates us to consider in a future work to improve the disambiguation process following this approach.

The tested measures were selected after comparing the PageRank algorithm (Brin and Page, 1998); for such measure, first, a directed graph was proposed but results were poor. Then, the graph was changed to an undirected representation, which was the better option, as show the results from Tables 1, 2, and 3. We think that Indegree is better because benefit of a large number of semantic relations, particularly of a densely connected graph. On the other hand, DH improved the results because, the context is not limited to a window size, wich determine the context range. Thus, a parser technique is used to describe the grammatical structure of the sentences, then the context of each word encountered in the corpus was extracted (see Section 3). This co-occurrence context-word is weighted according to their frequency in the corpus. We conclude that using the parser is better instead of a neighborhood around the ambiguous word.

The Table 4 shows a comparison with the works presented in the SemEval 2010 competition. The implemented system got a better performance than other systems, approximately 10% more on the precision and recall. Moreover, the obtained results are slightly low in comparison with the best one.

8 CONCLUSIONS AND FURTHER WORK

In this research, an approach for WSD on specific domain was presented. In such approach we have suggested a new method that uses the local and related context to retrieve second order vectors from WordNet to disambiguate combining both information. The experimental results comparing with SemEval 2010 showed promising results in precision and recall. As further work we think that better results could be gained using some techniques to extract key terms from an additional corpus to increase the size of the original corpus, as well as other semantic similarity measures to extract related words for an ambiguous word.

REFERENCES

Agirre, E., lopez de Lacalle, O., Fellbaum, C., Hsieh, S., Tesconi, M., Monachini, M., Vossen, P., and Segers, R. (2010). Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In *In Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010), Association for Computational Linguistics*.

Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *In Proc. of EACL, pages 3341*.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Computer Networks 30, 107117*.

Catherine, M. and Manning, C. (2008). Stanford typed dependencies manual. In *Technical report, Stanford University*.

Gale, W., Church, K., and Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. In *Comput. Human. 26, 415439*.

Iosif, E. and Potamianos, A. (2009). Unsupervised semantic similarity computation between terms using web documents. In *IEEE Transactions on Knowledge and Data Engineering Volume 22. no. 11, pp. 1637-1647*.

Kenneth, W. and Patrick, H. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics, vol. 16, no. 1, pp. 22-29*.

Navigli, R. (2009). Word sense disambiguation: A survey. In *A survey. ACM Computing Surveys, 41(2), Article 10*.

Navigli, R., Faralli, S., Soroa, A., Lopez de Lacalle, O., and Agirre, E. (2011). Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*.

Navigli, R. and Lapata, M. . (2007). Graph connectivity measures for unsupervised word sense disambiguation. In *In Veloso, M.M., ed.: IJCAI. (2007) 16831688*.

Patwardhan, S. and Ted, P. (2006). Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *In Proceedings of the Workshop on Making Sense of Sense at the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), pp. 18*.

Pedersen, T., Patwardhan, S., and Michelizzo, J. (2004). Wordnet::similarity - measuring the relatedness of concepts. In *In Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), pages 1024-1025*.

Reddy, S., Inumella, A., McCarthy, D., and Stevenson, M. (2010). Iiith: Domain specific word sense disambiguation. In *In Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 387-391, Uppsala, Sweden.*

Richard, J. (2004). From distributional to semantic similarity. In *Ph.D. Dissertation, University of Edinburgh*.

Schütze, H. and Pedersen, J. (1995). Information retrieval based on word senses. In *In Proceedings of SDAIR95 (Las Vegas, NV). 161175*.

Sinha, R. and Mihalcea, R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *In Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007), Irvine, CA, USA.*

Turney, P. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. In *Journal of Artificial Intelligence Research, 37:141188.*

Zhao, S. and Lin, D. (2004). A nearest-neighbor method for resolving pp-attachment ambiguity. In *In Proceedings of the IJCNLP-04.*