Enriched Page Rank for Multilingual Word Sense Disambiguation

Diego De Cao, Roberto Basili, Matteo Luciani, Francesco Mesiano, Riccardo Rossi

Dept. of Computer Science,
University of Roma Tor Vergata, Roma, Italy
{basili,decao}@info.uniroma2.it
{matteo.lcn,fra.mesiano,ricc.rossi}@gmail.com

Abstract. Word Sense Disambiguation (WSD) is an hard challenge especially for language different from english. Porting supervised models, that are state-of-art for english, on different languages is too much expensive. So unsupervised or semi-supervised WSD models are much more applicable to different languages. Graph-based methods, have been recently applied to linguistic knowledge bases, including unsupervised WSD. Although the achievable accuracy is rather high, the quality of the involved resources is de facto a crucial success factor. In this paper an adaptation of the PageRank algorithm proposed for WSD using distributional information is presented. This solution looks to preserve the achievable accuracy for the english language over a foreign language. An experimental analysis for the italian using standard benchmarks will be presented in the paper to support our hypothesis.

1 Background and Motivations

Lexical ambiguity is a fundamental aspect of natural language. Word Sense Disambiguation (WSD) investigates methods to automatically determine the intended sense of a word in a given context according to a predefined set of sense definitions, provided by a semantic lexicon. Intuitively, WSD can be usefully exploited in a variety of NLP and Information Retrieval tasks such as *ad hoc retrieval* [1, 2] or Question Answering [3]. However controversial results have been often obtained, as for example the study on text classification reported in [4]. The impact of WSD on IR tasks is still an open issue and large scale assessment is needed. For this reason, unsupervised approaches to inductive WSD are appealing.

It has been more recently that graph-based methods for knowledge-based WSD have gained much attention in the NLP community ([5–7]). In these methods a graph representation for senses (nodes) and relation (edges) is first built. In [7], a comparative analysis of different graph-based models based on PageRank model [8] over two well known WSD benchmarks is reported. A special emphasis for the resulting computational efficiency is also posed there. In particular, a variant called *Personalized PageRank* (PPR) is proposed in [7]. This variant tries to trade-off between the amount of the employed lexical information and the overall efficiency. In synthesis, along the ideas of the Topic sensitive PageRank [9], PPR suggests that a proper initialization of the teleporting vector $\boldsymbol{p}$ suitably captures the context information useful to drive the random surfer PageRank model over the graph to converge towards the proper senses in fewer steps. In [10] we present a model to extend the PPR trough distributional evidence improving the overall *PPR* performances over the English language. In this paper we discuss the applicability of the extension of *PPR* algorithm to Italian language.

The key idea is to exploit an externally acquired semantic space to expand the incoming sentence σ into a set of *novel* terms, different but *semantically related* with the words in σ. In analogy with topic-driven PageRank, the use of these words as a seed for the iterative algorithm is expected to amplify the effect of local information (i.e. σ) onto the recursive propagation across the lexical network: the interplay of the global information provided by the whole lexical network with the local information characterizing the initialization lexicon is expected to maximize their independent effect.

More formally, let the matrix $W_k := U_k S_k$ be the matrix that represents the lexicon in the k-dimensional Latent Semantic Analysis (LSA) [11] space. Given an input sentence σ, a vector representation $\overrightarrow{w_i}$ for each term w_i in σ is made available. The corresponding representation of the sentence can be thus computed as the linear combination through the original $tf \cdot idf$ scores of the corresponding $\overrightarrow{w_i}$: this provides always an unique representation $\overrightarrow{\sigma}$ for the sentence. $\overrightarrow{\sigma}$ locates the sentence in the LSA space and the set of terms that are *semantically related* to the sentence σ can be easily found in the neighborhood. A lower bound can be imposed on the cosine similarity scores over the vocabulary to compute the lexical expansion of σ, i.e. the set of terms that are enough similar to $\overrightarrow{\sigma}$ in the k dimensional space. Let D be the vocabulary of all terms, we define as the lexical expansion $T(\sigma) \subset D$ of $\overrightarrow{\sigma}$ as follows:

$$T(\sigma) = \{w_j \in D : sim(\overrightarrow{w_j}, \overrightarrow{\sigma}) > \tau\} \tag{1}$$

where τ represents a real-valued threshold in the set $[0, 1)$. In order to improve precision it is also possible to impose a limit on the cardinality of $T(\sigma)$ and discard terms characterized by lower similarity factors.

Finally, the later steps of the PPR methods remain unchanged, and the PageRank works over the corresponding graph.

2 Empirical Evaluation

The evaluation of the proposed model was focused to evaluate the applicability of the *Extended PPR* to the Italian language. This will be done also comparatively with the state of the art of unsupervised systems over a consolidated benchmark, Evalita 2007 for the Italian language. Concerning to the distributional approach the Italian Web as Corpus[1] (about 1800K web pages) is used with about 150k words. The corpus is processed with the TreeTagger[2] to extract the part of speech for every words. Then a dimensionality reduction factor of $k = 100$ is adopted to perform the LSA space. For the italian language the ItalWordNet [12] resource is adopted. Two different approaches have been employed for the process of Word Sense Disambiguation:

- Sentence based approach: the process of LSA expansion and disambiguation is performed for every single sentence of the dataset

[1] http://wacky.sslmit.unibo.it/

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

- Document based approach: the process of LSA expansion and disambiguation is performed for every document of the dataset. In this approach we used a policy of ”one sense per discourse”.

The Evalita ’07 all-words task[3] consists of about 4700 words to disambiguate. Due the novel configuration of the distributional space that came out using the Italian corpus, the damping factor, the number of iterations and the number of words for the LSA expansion have been re-estimated. The Table 1 reports the result at different parameters for sentence and document based approaches over the Evalita ’07 test data. For each parameter the columns in tables show Precision, Recall and F-Measure for the PPR and PPRw2w respectively.

	Parameters			PPR			w2w		
	WN	LSA	Iter.	Prec	Rec	F1	Prec	Rec	F1
Sentence	16	80	25	57.1	46.1	**51.0**	57.6	46.5	**51.4**
	16	0	25	56.6	45.7	50.6	56.5	45.3	50.3
Document	16	800	25	—	—	—	58.5	47.3	**52.3**
	16	1000	25	58.4	47.3	**52.3**	—	—	—
	16	0	25	57.1	46.1	51.0	57.6	46.5	51.4

Table 1. Accuracy of the LSA-based expansion PageRank model, as compared with thePPR standard and word oriented ($w2w$) versions of the personalized PageRank over the Evalita 2007 datasets. 16 refers to the resource of ItalWordNet 1.6

Algorithm	Precision	Recall	F1	Attempted
UKB_LSA_Sent	57.6	46.5	51.4	79.3
UKB_LSA_Doc	58.5	47.3	**52.3**	80.0
UKB	57.0	46.1	51.0	79.8
JIGSAW	56.00	41.40	47.60	73.95
First Sense Baseline	66.9	66.9	66.9	100

Table 2. Accuracy of the different tools over the Evalita 2007

We adopted fixed limits for LSA expansion where values from 20 up to 1000 terms have been tested. The good scores obtained on the [10] suggested that a number of iterations lower than 30 is in general enough to get good accuracy levels: 25 iterations, instead of 30, have been judged adequate. Finally, on average, the total number of lexical items in the expanded sentence $T(\sigma)$ includes about 40% of nouns, 30% of verbs, 20% of adjectives and 10% of adverbs. As a confirmation of the outcome in [7], the word-by-word model achieves better results. Interestingly, almost on every type of graph and for every approach (sentence or word oriented) the LSA-based method outperforms the original UKB.

Table 2 reports Precision, Recall and F1 scores of the different systems as obtained over the test Evalita ’07 data. The best F1 scores between any pair are emphasized in bold, to comparatively asses the results. Results confirms that the impact of the topical

[3] http://evalita.fbk.eu/

information provided by the LSA expansion of the sentence is beneficial for a better use of the lexical graph. An even more interesting outcome is that the improvement implied by the proposed LSA method on the sentence oriented model (i.e. the standard PPR method of [7]) is higher, so that the difference between the performances of the $PPRw2w$ model are no longer strikingly better than the PPR one. As shown in Table 2 our method outperforms the JIGSAW system of over 9.87% in the F-Measure. Moreover the good accuracy reachable by the document-based approach is also very interesting as for the higher time efficiency of this approach with respect to the sentence-based one. As a matter of fact with the first approach the system has been run only sixteen times instead of hundreds times when the sentence-based approach is employed. Furthermore the lower execution times suggest the applicability of the system to different Information Retrieval scenario, such as the Question Answering or the Cross Language Information Retrieval (CLIR). CLIR is a challenging task and the existence of aligned lexical database, such as MultiWordnet[4] that is aligned to the English WordNet version, opens an interesting perspective of using word senses as anchor to search in different language.

References

1. Krovetz, H.: Homonymy and polysemy in information retrieval. In: Proceedings of the 35th ACL '09. (1997)
2. bum Kim, S., cheol Seo, H., chang Rim, H.: Information retrieval using word senses: root sense tagging approach. In: In SIGIR 2004. (2004) 258–265
3. Beale, S., Lavoie, B., McShane, M., Nirenburg, S., Korelsky, T.: Question answering using ontological semantics. In: TextMean '04, ACL (2004) 41–48
4. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In: Proc. of the European Conf. on IR, ECIR, New York, USA (2004) 181–196
5. Sinha, R., Mihalcea, R.: Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In: IEEE ICSC 2007. (2007)
6. Navigli, R., Lapata, M.: Graph connectivity measures for unsupervised word sense disambiguation. In: Proceedings of IJCAI'07. (2007)
7. Agirre, E., Soroa, A.: Personalizing pagerank for word sense disambiguation. In: Proceedings of the 12th conference of EACL '09, Athens, Greece (March 30 - April 3 2009)
8. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30**(1–7) (1998) 107–117
9. Haveliwala, T.H.: Topic-sensitive pagerank. In: Proc. of 11th Int. Conf. on World Wide Web, New York, USA, ACM (2002)
10. De Cao, D., Basili, R., Luciani, M., Mesiano, F., Rossi, R.: Robust and efficient page rank for word sense disambiguation. In ACL, ed.: Proceeding of TextGraphs-5. (2010)
11. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review **104** (1997)
12. A Alonge Roventini, A., Bertagna, F., Calzolari, N., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A.: Italwordnet: Building a large semantic database for the automatic treatment of italian. In: Linguistica Computazionale, IEPI, Pisa-Roma. (2003)

[4] http://multiwordnet.fbk.eu/