PageRank-based Word Sense Induction within Web Search Results Clustering

José G. Moreno, Gaël Dias

PageRank-based Word Sense Induction within Web Search Results Clustering

Jose G. Moreno
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
jose.moreno@unicaen.fr

Gaël Dias
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
gael.dias@unicaen.fr

ABSTRACT

Word Sense Induction is an open problem in Natural Language Processing. Many recent works have been addressing this problem with a wide spectrum of strategies based on content analysis. In this paper, we present a sense induction strategy exclusively based on link analysis over the Web. In particular, we explore the idea that the main different senses of a given word share similar linking properties and can be found by performing clustering with link-based similarity metrics. The evaluation results show that PageRank-based sense induction achieves interesting results when compared to state-of-the-art content-based algorithms in the context of Web Search Results Clustering.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval—*clustering*

General Terms

Algorithms, Experimentation

Keywords

Word Sense Indiction, Web Links, PageRank Clustering

1. INTRODUCTION

Word Sense Induction (WSI) is an open problem in Natural Language Processing (NLP), which has fostered a great deal of attention in the past few years. Indeed, many recent works have been addressing this problem by analysing Web contents and exploring interesting ideas to extract knowledge from external resources. One important work is proposed by [2] whose evidence increased performance within Web Search Results Clustering (SRC) when WSI is performed over the Google Web1T corpus.

In this paper, we present a WSI strategy exclusively based on link analysis over Web collections. The underlying idea is simple and grounded on the following hypothesis: *word senses are distributed over the Web in the same way Web pages are linked together*. In other words, Web pages containing the same word meaning should share some similar link-based values. This hypothesis supposes that (1) senses are separated by linking importance of the Web and (2) Web domains provide a unique meaning of a given word, thus extrapolating the "one sense per discourse" paradigm defined by [3]. So, if both factors are true, word senses should be found by performing clustering over link-based similarity metrics where one cluster represents a unique sense.

2. WORD SENSE INDUCTION

Automatically discovering word senses is a challenging research topic. Recent works have been concentrating on the evaluation of WSI through Web search results clustering [2]. In particular, task 11 of the SemEval13 challenge is dedicated to this issue [6]. The main problem consists of cluster Web pages with similar senses from a given list of Web search results. All evaluated strategies use Web snippet content and/or external resources such as Wikipedia. In this paper, we propose to study the importance of link analysis for WSI. It is important to notice that our approach does not take into account any content and uniquely relies on linking relations between returned Web pages. To the best of our knowledge, this is the first attempt to solve WSI without content information. As such, we aim to propose another perspective to solve an important issue in NLP and IR.

3. PAGERANK-BASED SENSE INDUCTION

PageRank-based clustering has proved to be a useful strategy for hypertext document clustering [1]. In this paper, we adapt these ideas to SRC as clustering is run over the sub-collection returned by the search engine and not over the entire Web collection. As a consequence, we propose to use the Jensen-Shannon Divergence metric to calculate similarities between hypertext documents[1]. Let us define $D_q = \{d_q^1, d_q^2, .., d_q^{n_q}\}$ as a list of Web results related with the query q, $PR_q = \{pr_q^1, pr_q^2, .., pr_q^{n_q}\}$ as the corresponding and known list of PageRanks values corresponding to the D_q hypertext documents. To calculate the kernel values between d_q^i and d_q^j, we use the Jensen-Shannon kernel proposed by [4]: $k_{JS}(d_q^i, d_q^j) = ln2 - JS(d_q^i, d_q^j)$, where the $JS(d_q^i, d_q^j)$ value is defined under the hypothesis that each hypertext document has a probability distribution with two states corresponding to be selected by a random walk or not.

[1] [1] discarded this option as it was computationally expensive in their reserach work over the entire Web.

For the first state, we consider the PageRank value (pr_q^l) as a probability value for the hypertext document l to be selected and $1-pr_q^l$ to not be. Given the huge size of the Web, we fairly assume that $1-pr_q^l$ is always near to one and as consequence $ln(1-pr_q^l)$ can be approximated to zero. Given this, we can formulate the JS divergence between two Web pages as is shown in Equation 1

$$JS(d_q^i, d_q^j) = \frac{1}{2}\left[pr_q^i ln\left(\frac{2*pr_q^i}{pr_q^i+pr_q^j}\right) + pr_q^j ln\left(\frac{2*pr_q^j}{pr_q^i+pr_q^j}\right)\right] \quad (1)$$

A final normalization step is performed to ensure 1s in the diagonal of the Kernel Matrix. As a clustering algorithm, we have chosen two alternatives: (1) the Spectral Clustering Algorithm[2] (PRSC-k) and (2) equal size partitions ordered by PageRank (PRSim-k). Each output cluster is considered as one unique sense and evaluated as it.

4. EXPERIMENTAL SETUP

Dataset In our experiments the SemEval13 Word Sense Induction dataset was used. A further description can be found in [6]. In brief, it is composed of 100 queries extracted from AOL query log dataset which have a corresponding Wikipedia disambiguation page. Each query has 64 Web results classified in one of the senses proposed in the Wikipedia article. However, the Web results does not include the PageRank values. For that, we have used the Hyperlink Graph publicly available in [5]. Each Web result is reduced to a Pay-Level-Domain Graph and a PageRank value is assigned after calculating all of them for the entire PLD Graph. The HyperLink Graph is composed of more than 43 million PLD values and less than 1.3% of the URLs of the SemEval13 dataset were not found. For these cases, the lowest PageRank value was assigned to avoid zero values. To evaluate the cluster quality, we selected the same SemEval13 metrics: F_1-measure (F1), RandIndex (RI), Adjusted RandIndex (ARI) and Jaccard coefficient (J).

Baselines As a simple baseline, two versions of a Random clustering algorithm were examined. Additionally, more competitive baselines were also implemented; one based on the well-known Latent Dirichlet Allocation (LDA) technique and the other one a tf-idf representation combined with the Spectral Clustering algorithm (TextSC). All parameters were selected to guarantee the best performance from each algorithm. Over these two, we explored the same number of clusters as we did with our algorithm to adequately analyse the performance of the non-content and content-based strategies.

5. RESULTS

The results for different k values of the ARI metric are presented in Figure 1. Note that we have included the six different algorithms. Consistently, ARI gives a score near to zero to both random-based strategies. PRSC-k and TextSC behave similarly for ARI, suggesting that no significant differences can be observed when the Spectral Clustering algorithm is applied over content-based or non-content-based similarities. Hypothetically, our achieved position of PRSC-k algorithm in SemEval13 WSI challenge is presented in Table 1. Note that our non-content-based algorithm is a competitive solution for WSI in terms of the analyzed metrics.

[2]Implemented in SciKit Learn tool http://scikit-learn.org/ [Last access: 11/06/2014].

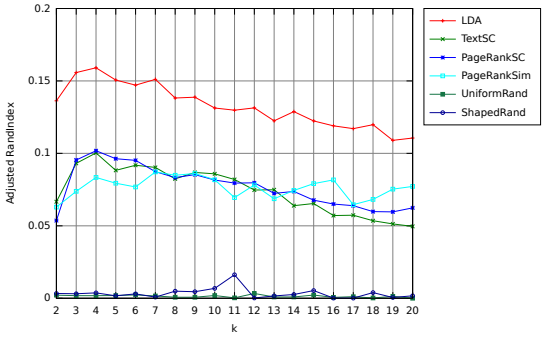

Figure 1: ARI values for content- and non-content-based algorithms. Cluster size varing from 2...20.

Algorithm	F1	RI	ARI	J
PRSC-5	0.5997 5^{th}	0.5782 5^{th}	0.0963 3^{rd}	0.2693 9^{th}
PRSC-10	0.6367 4^{th}	0.5688 5^{th}	0.0816 3^{rd}	0.2406 9^{th}
PRSim-5	0.6089 4^{th}	0.6048 3^{rd}	0.0794 3^{rd}	0.2098 9^{th}
PRSim-10	0.6456 4^{th}	0.6237 3^{rd}	0.0818 3^{rd}	0.1593 9^{th}

Table 1: Perfomance and hipotethical position (*in italics*) in the SemEval13 WSI task.

6. CONCLUSIONS

In this paper, we have presented a WSI algorithm based on PageRank clustering. Results show that non-content-based clustering algorithms can achieve competitive results when compared with content-based. In consequence, the PageRank clustering algorithm allows us to capture each sense in each cluster. As future work, we propose that the combination between content and non-content-based strategies could allow the improvement of the overall performance of the WSI systems.

7. REFERENCES

[1] K. Avrachenkov, V. Dobrynin, D. Nemirovsky, S. Pham, and E. Smirnova. Pagerank based clustering of hypertext document collections. In *Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 873–874, 2008.

[2] A. Di Marco and R. Navigli. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(4):709–754, 2013.

[3] W. Gale, K. Church, and D. Yarowsky. One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language (HLT)*, pages 233–237, 1992.

[4] A. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo. Nonextensive information theoretic kernels on measures. *The Journal of Machine Learning Research*, 10:935–975, 2009.

[5] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer. Graph structure in the web - revisited. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 427–432, 2014.

[6] R. Navigli and D. Vannella. Semeval-2013 task 11: Word sense induction & disambiguation within an end-user application. In *Proceedings of the International Workshop on Semantic Evaluation (SEMEVAL)*, pages 1–9, 2013.