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 Induction, 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 [1] who 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. 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.

Automatically discovering word senses is a challenging research topic. Recent works have been concentrating on the evaluation of WSI through Web search results clustering [1]. In particular, task 11 of the SemEval13 challenge is dedicated to this issue [2] and consists of clustering Web pages with similar senses from a given list of Web search results. Within this context, 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.

2. PAGERANK-BASED STRATEGY

PageRank-based clustering has proved to be a useful strategy for hypertext document clustering [3]. In this paper, we adapt these ideas to SRC as clustering is run over the subcollection 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 $D_q = \{d_q^1, d_q^2, .., d_q^n\}$ the list of Web results related to the query q and $PR_q = \{pr_q^1, pr_q^2, .., pr_q^n\}$ the previously computed list of PageRank values corresponding to each of the hypertext documents in D_q. To calculate the kernel values between two Web pages d_q^i and d_q^j, we use the Jensen-Shannon (JS) kernel proposed by [4] such that $k_{JS}(d_q^i, d_q^j) = ln2 - JS(d_q^i, d_q^j)$, where the $JS(d_q^i, d_q^j)$ value is defined under the hypothesis that each hypertext document is associated to a probability distribution with two states. The first state value corresponds to the PageRank value pr_q^l

[1][3] discarded this option as it was computationally expensive in their research work over the entire Web.

978-1-4799-5569-5/14/$31.00

i.e. the probability value for the hypertext document d_p^l to be selected over a random walk. Correspondingly, the second state value is defined as $1-pr_q^l$, which corresponds to the non-selection of the d_p^l. 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 Jensen-Shannon divergence between two Web pages as in Equation 1. Note that a final normalization step is performed to ensure values equals to 1 in the diagonal of the kernel matrix.

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

With respect to clustering, we chose two alternatives: Spectral Clustering[2] (PRSC-k) and Equal size partitions ordered by PageRank values (PRSim-k). Finally, each output cluster is considered as one unique sense and evaluated as such to the reference.

3. EXPERIMENTAL SETUP

In our experiments, the SemEval13 Word Sense Induction dataset was used. A further description can be found in [2]. In brief, it is composed of 100 queries extracted from the AOL query log dataset, each one having a corresponding Wikipedia disambiguation page. Each query is associated to 64 Web search results classified in one of the senses proposed in the Wikipedia article. However, Web search results do not include PageRank values. As a consequence, we used the publicly available Hyperlink Graph from [5]. Each Web search result is reduced to a Pay-Level-Domain (PLD) Graph and a PageRank value is assigned to it after calculating all of them for the entire PLD Graph. In particular, 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 in it. For these cases, the lowest PageRank value was assigned to avoid zero values. To evaluate cluster quality, we selected the metrics proposed in SemEval13 i.e. F_1-measure (F1), RandIndex (RI), Adjusted RandIndex (ARI) and Jaccard coefficient (J).

As simple baselines, two versions of a Random clustering algorithm were examinated. Additionally, more competitive baselines were implemented: (1) Latent Dirichlet Allocation (LDA) technique and (2) a tf-idf representation combined with the Spectral Clustering algorithm (TextSC). All parameters were selected to guarantee the best performance for each algorithm and we explored the same number of clusters to adequately analyse the performance of the non-content and content-based strategies.

4. RESULTS

The results for different k cluster 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 strategies. PRSC-k and TextSC behave similarly for ARI, suggesting that no significant difference can be observed when the Spectral Clustering algorithm is applied over content-based or non-content-based similarities. Hypothetically, the position of PRSC-k 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, although it is outperformed by LDA.

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

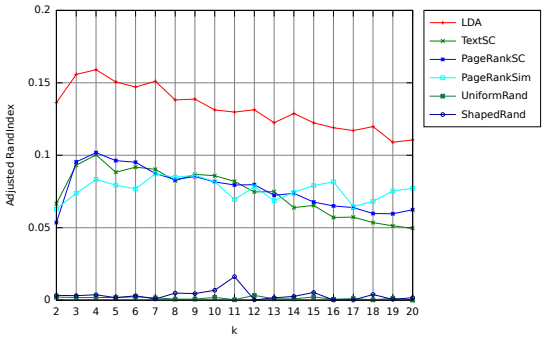

Figure 1: ARI values for content- and non-content-based algorithms. Cluster size varying from 2..20.

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

Table 1: Overall perfomance and hypothetical position (*in italics*) in task 11 of SemEval13.

5. CONCLUSIONS

In this paper, we presented a WSI algorithm based on PageRank clustering. Results show that non-content-based strategies can achieve competitive results when compared to "simple" content-based ones and envision that the combination of both strategies may allow improvements for WSI.

6. REFERENCES

[1] A. Di Marco and R. Navigli, "Clustering and diversifying web search results with graph-based word sense induction," *Computational Linguistics*, vol. 39, no. 4, pp. 709–754, 2013.

[2] 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)*, pp. 1–9, 2013.

[3] 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)*, pp. 873–874, 2008.

[4] A. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo, "Nonextensive information theoretic kernels on measures," *The Journal of Machine Learning Research*, vol. 10, pp. 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)*, pp. 427–432, 2014.