# Multi-Objective Word Sense Induction using Content and Interlink Connections

Sudipta Acharya[1], Asif Ekbal[1], Sriparna Saha[1], Prabhakaran Santhanam[1],
Jose G. Moreno[2], and Gaël Dias[3]

[1] Indian Institute of Technology Patna
[2] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
[3] University of Caen Lower Normandy, France

**Abstract.** In this paper, we propose a multi-objective optimization based clustering approach to address the word sense induction problem by leveraging the advantages of document-content and their structures in the Web. Recent works attempt to tackle this problem from the perspective of content analysis framework. However, in this paper, we show that contents and hyperlinks existing in the Web are important and complementary sources of information. Our strategy is based on the adaptation of a simulated annealing algorithm to take into account second-order similarity measures as well as structural information obtained with a pageRank based similarity kernel. Exhaustive results on the benchmark datasets show that our proposed approach attains better accuracy compared to the content based or hyperlink strategy encouraging the combination of these sources.

## 1 Introduction

Word Sense Induction (WSI) is a crucial problem in Natural Language Processing (NLP), which has drawn significant attention to the researchers during the past few years. WSI concerns the automatic identification of the senses of a known word, which is an expected capability of modern information retrieval systems. In recent times, there are few research works that address this problem by analyzing Web contents and exploring interesting ideas to extract knowledge from external resources. One important work is proposed in [6], which shows that increased performance may be obtained for Web Search Results Clustering (SRC) when word similarities are calculated over the Google Web1T corpus. The authors propose a comparative evaluation of WSI systems by the use of an end-user application such as SRC. The key idea behind SRC system is to return some meaningful labeled clusters from a set of snippets retrieved from a search engine for a given query. So far, most of the works have focused on the discovery of relevant and informative clusters [4] in which the results are organized by topics in such a way as WSI.

In this paper, we present a strategy for WSI based on content and link analyses over Web collections through the use of a Multi-Objective Optimization (MOO) technique [5]. The underlying idea is grounded on the hypothesis that word senses are related to the distribution of words on Web pages and the way that they are linked together. In other words, Web pages containing the same word meaning should share some similar content-based and link-based values. This study supposes that (1) word distribution is different for each sense, (2) links over the Web express knowledge complementary to content and (3) web domains provide an unique meaning of a word, thus extrapolating the "one sense per discourse" paradigm [7].

Our proposal addresses the WSI problem using a MOO framework that has a different perspective compared to Single Objective Optimization (SOO). In SOO, we concentrate in optimizing only a single objective function, whereas MOO deals with the issue of simultaneously optimizing more than one objective function. Here, we first pose the problem of WSI within the framework of MOO, and thereafter solve this using a simulated annealing based MOO technique called AMOSA [3]. Specifically, we are interested in optimizing the similarity of a cluster of documents in terms of both their content and interlink similarity. The combination of these sources is not straightforward, and this calls for the use of MOO techniques.

The main contributions of this paper are summarized as follows. First, we propose a new MOO algorithm for WSI which automatically determines the number of senses. Next, we propose an evaluation of alternative sources as a combination to improve existing solutions to the WSI problem. Evaluation results show that our MOO based clustering algorithm performs better compared to the hyperlink-based techniques, and very closely compared to approaches with strong content-based solutions when evaluated using WSI measures.

## 2 Related Works

As far as our knowledge goes, no existing methodology for WSI uses the hyperlink information and content information in an unified setting. The popular techniques either use one or the other sources of information. *Content-based* techniques such as [6] are based on the use of word distributions over the huge collections of n-grams mapped to a graph where the nodes are the words (sense candidates) and the arcs are calculated based on the frequencies in which two words are found together. Once the graph is built the words are grouped together based on simple graph patterns such as curvature clustering or balanced maximum spanning tree clustering, where each obtained cluster represents a sense. Similarly, [8] proposes the use of extra frequency information extracted from Wikipedia and forms groups using a variation of the well known latent dirichlet allocation (LDA) algorithm. Each topic obtained by LDA is considered as a cluster. In contrast, *hyperlink-based* techniques are quite rare. In [11] authors have proposed a technique exclusively based on hyperlink information. There are works where similarity between documents is calculated using a Jensen-Shannon kernel [9] and then clustering is performed by the use of a classical spectral clustering algorithm. Each cluster represents a sense in the solution. Both approaches manage to adequately discover the word senses in the documents. However, reported results show a superiority of content-based techniques over hyperlink-based techniques.

In the field of application of MOO, as far as we know, within text applications, [12] is the first work, which formulates text clustering as a MOO problem. In particular, they express desired properties of frequent itemset clustering in terms of multiple conflicting objective functions. The optimization is solved by a genetic algorithm and the result is a set of Pareto-optimal solutions. But, towards solution of WSI problem, according to best of our knowledge, this is the very first attempt that utilises MOO based clustering approach.

## 3 Combining Content and Hyperlink Approaches with a MOO

Most of the existing SRC techniques are based on a single criterion which reflects a single measure of goodness of a partitioning. However, one single source of information may not fit all problems. In particular, [2] have shown the utility of hyperlink information to cluster Web documents, whereas [11] have shown their applicability to the WSI problem. Manifestly, both SRC and WSI have been addressed by the analysis of document contents. Moreover, an effective combination of these two has not been yet proposed. Hence, it may become necessary to simultaneously optimize several cluster quality measures which can capture different content-based or hyperlink-based characteristics. In order to achieve this, MOO can be an ideal platform and therefore we pose the problem of finding word senses within this framework. Therefore, the application of sophisticated MOO techniques seems appropriate and natural.

**Content Compactness based on SCP/PMI Measure:** This type of indices measures the proximity among the various elements of the cluster. One of the commonly used measures for compactness is the variance. Documents are kept together if the distribution of words is similar. This measurement is based on either Symmetric Conditional Probability ($SCP$) or Pointwise Mutual Information ($PMI$), defined on section 4.2.

**Content Separability based on SCP/PMI Measure:** This particular type of indices is used in order to differentiate between two clusters. Distance between two cluster centroids is a commonly used measure of separability. This measure is easy to compute and can detect hyperspherical-shaped clusters well. Word-distribution between senses must be as different as possible. This measurement is also based on computing either $SCP$ or $PMI$.

**Hyperlink Compactness:** Well-connected Web documents must belong to the same sense cluster.

**Hyperlink Separability:** The number of interlinks between senses must be as small as possible.

## 4 MOO based Clustering for WSI

In order to perform MOO based clustering we adapt Archived multi-objective simulated annealing (AMOSA) [3] as the underlying optimization strategy. It incorporates the concept of an archive where the non-dominated solutions seen so far are stored. Steps of the proposed approach are described in the following sections. For better understanding we also show the various steps of our proposed algorithm in Figure 1.

### 4.1 Archive Initialization

As we follow an endogenous approach, only the information returned by a search engine is used. In particular, we only deal with web snippets and each one is represented as a word feature vector. So, our proposed clustering technique starts its execution after initializing the archive with some random solutions as archive members. Here, a particular solution refers to a complete assignment of web snippets (or data points) in several clusters. So, the first step is to represent a solution compatible with AMOSA, which represents each individual solution as a string. In order to encode the clustering problem in the form of a string, a center-based representation is used. Note that the use of a string representation facilitates the definition of individuals and mutation functions [3].

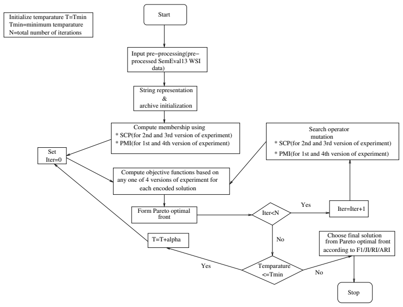
Initialize temparature T=Tmin
Tmin=minimum temparature
N=total number of iterations
Start
Input pre−processing(pre−processed SemEval13 WSI data)
String representation & archive initialization
Compute membership using
* SCP(for 2nd and 3rd version of experiment)
* PMI(for 1st and 4th version of experiment)
Set Iter=0
Search operator mutation
* SCP(for 2nd and 3rd version of experiment)
* PMI(for 1st and 4th version of experiment)
Compute objective functions based on any one of 4 versions of experiment for each encoded solution
Form Pareto optimal front
Iter<N
Yes
Iter=Iter+1
No
T=T+alpha
Choose final solution from Pareto optimal front according to F1/JI/RI/ARI
Yes
Temparature <=Tmin
No
Stop

**Fig. 1.** Flowchart of proposed methodology

Let us assume that the archive member $i$ represents the centroids of $K_i$ clusters and the number of tokens in a centroid is $p^4$, then the archive member has length $l_i$ where $l_i = p \times K_i$. To initialize the number of centroids $K_i$ encoded in the string $i$, a random value between 2 and $K_{max}$ is chosen and each of the $K_i$ centroids is initialized by randomly generated token from the vocabulary.

### 4.2 Content-based Similarity Measure

In order to compute the similarity between two Web documents (word vectors) we have used two well known content-based similarity metrics, the $SCP(W_1, W_2) = \frac{P(W_1,W_2)^2}{P(W_1) \times P(W_2)}$ and $PMI(W_1, W_2) = \log(\frac{P(W_1,W_2)}{P(W_1) \times P(W_2)})$.

We have computed the $SCP$ and $PMI$ values between each pair of words in the global vocabulary ($V$), i.e., the set of different tokens in the list of all Web documents.

### 4.3 Hyperlink-based Similarity Measure

Given a collection of Web documents relevant to a sense, we calculate their corresponding pagerank values $pr_{d_i}$. Similarity between documents is calculated through a kernel function between pagerank values. Specifically, we use the Jensen-Shannon kernel proposed by [9]: $k_{JS}(d_i, d_j) = ln2 - JS(d_q^i, d_q^j)$, where the $JS(d_i, d_j)$ value is defined under the hypothesis that each hypertext document has a probability distribution with two states: whether or not they are selected by a random walk. Following the proposal for pagerank similarities defined by [11], we calculate the similarity values between two Web pages as shown in Equation 1

$$JS(d_i, d_j) = \frac{1}{2}\left[p^r_{d_i} ln\left(\frac{2 * p^r_{d_i}}{p^r_{d_i} + p^r_{d_j}}\right) + p^r_{d_j} ln\left(\frac{2 * p^r_{d_j}}{p^r_{d_i} + p^r_{d_j}}\right)\right]. \tag{1}$$

These similarity measures (two content-based and one hyperlink-based) are used in our proposed MOO based clustering framework.

### 4.4 Assignment of Web-snippets and Objective Function Calculations

After initializing the archive the first step concerns the assignment of $n$ word vectors or points (where $n$ is the total number of Web snippets in a particular query) to different clusters. This assignment can be done using any one of content-based similarity measurement techniques ($SCP$ or $PMI$). In the second step, we compute two cluster

[4] A centroid is represented by a $p$ word feature vector $(w_k^1, w_k^2, w_k^3, \ldots, w_k^p)$.

quality measures, cluster compactness and separation, by varying similarity computation (either content-based or hyperlink-based) and use them as objective functions of the string. Thereafter we simultaneously optimize these objective functions using the search capability of AMOSA.

**Assignment of Web-snippets and Updation of Centroids** In this technique, the assignment of points to different clusters is done based on the content similarity measurement between that point and different cluster centroids. Each Web document is assigned to that cluster center with respect to which it has the maximum similarity measure. In particular, any point $j$ is assigned to a cluster $t$ whose centroid has the maximum similarity to $j$ using:

$$t = argmax_{k=1,...K} S(\overline{x}_j, \overline{m}_{\pi_k}). \tag{2}$$

$K$ denotes the total number of clusters, $\overline{x}_j$ is the $j^{th}$ point (or Web document), $\overline{m}_{\pi_k}$ is the centroid of the $k^{th}$ cluster $\pi_k$ and $S(\overline{x}_j, \overline{m}_{\pi_k})$ denotes similarity measurement between the point $\overline{x}_j$ and cluster centroid $\overline{m}_{\pi_k}$.

One possible way to compute the similarity between two word vectors is defined by Equation 3, which is inspired by [1] [5].

$$S(d_i, d_j) = \sum_{k=1}^{\|d_i\|} \sum_{l=1}^{\|d_j\|} SCP(W_{ik}, W_{jl}) \tag{3}$$

Here $\|d_i\|$ and $\|d_j\|$ respectively denote, total number of words in word vectors $d_i$ and $d_j$. After assigning all Web snippets to different clusters, the cluster centroids encoded in that string are updated. For each cluster, $p$ number of words from global vocabulary which are most similar to other documents of that particular cluster are chosen to form new centroid of that cluster.

**Objective Functions** In order to compute the goodness of the partitioning encoded in a particular string, cluster compactness and separability are usually used as the objective functions in MOO clustering. The objective functions quantify two intrinsic properties of the partitioning. First, compactness is defined in Equation 4 and it is maximized.

$$COM = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} S(x_i, m_{\pi_k}) \tag{4}$$

Here $m_{\pi_k}$ is the cluster centroid of the $k^{th}$ cluster consisting of $p$ words $(w_1^{\pi_k}, ....., w_p^{\pi_k})$, $K$ is the number of clusters encoded in that particular string and $S(x_i, m_{\pi_k})$ value is computed using Equation 3. In Equation 3, SCP can be replaced by PMI. Also cluster compactness can be measured using hyperlink similarity as given in Equation 1. The compactness using hyperlink similarity is computed by the following equation:

$$COM = \sum_{k=1}^{K} \frac{\sum_{x_i, x_j \in \pi_k} JS(x_i, x_j)}{|\pi_k|} \tag{5}$$

Hence, total three different versions of cluster compactness can be computed by varying the similarity measures. Note that if words in a particular cluster are very similar to the cluster centroid and documents are highly interconnected then the corresponding $COM$ value would be maximized. Also for hyperlink similarity based compactness if all the documents of any particular cluster are highly interconnected to each other then also corresponding $COM$ gets maximized. Here our target is to form good clusters whose compactness in terms of similarity should be maximum.

[5] SCP could be replaced by PMI.

The second objective function is the cluster separation which measures the dissimilarity between two given clusters. Purpose of any clustering algorithm is to obtain compact similar typed clusters which are dissimilar to each other. Here we have computed the summation of similarities between different pairs of cluster centers and then minimized this value just to produce well-separated clusters. The separation is defined in Equation 6.

$$SEP = \sum_{k=1}^{K} \sum_{o=k+1}^{K} S(m_{\pi_k}, m_{\pi_o}) \tag{6}$$

Here $m_{\pi_k}$ and $m_{\pi_o}$ are the centroids of the clusters $\pi_k$ and $\pi_o$, respectively. $S(m_{\pi_k}, m_{\pi_o})$ value is computed using Equation 3.

Similar to compactness, SCP or PMI based similarity measure can be used to compute separatibility. The process to compute $SEP$ value using hyperlink-based similarity measure for a string is given in Equation 7.

$$SEP = \sum_{k=1}^{K} \sum_{o=k+1}^{K} \forall_{x_i \in k, x_j \in o} min(JS(x_i, x_j)) \tag{7}$$

It shows that the sum of maximum distance (i.e., minimum similarity) between the documents of all possible pairs of clusters in a string is represented as the separability measure. Minimizing this value represents well separated clusters.

Therefore, similar to compactness, separation $SEP$ can be calculated in three different ways by varying the similarity measures. Out of total six compactness and separation based objectives any combination of them can be used. These objective functions are maximized using the search capability of AMOSA.

### 4.5 Search Operators

As mentioned earlier, the proposed clustering technique uses a multi-objective simulated annealing based approach as the underlying optimization strategy. As a simulated annealing step, we have introduced three types of mutation operations as used in [1]. These mutation operations can update, increase or decrease the size of a string. During smilarity measurement in mutation operations either SCP or PMI similarity matrix is used. In order to generate a new string any one of the above-mentioned mutation types is applied to each string with equal probability.

## 5 Experimental Setup

**Dataset:** In our experiments the SemEval13 Word Sense Induction dataset [13] was used. In brief, it is composed of 100 queries extracted from AOL query log dataset which has 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 do not include the PageRank values. For that, we use the Hyperlink Graph publicly available in [10]. Each Web result is reduced to a Pay-Level-Domain (PLD) 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 (JI).

**Baselines:** As baselines, we use the well-known Latent Dirichlet Allocation (LDA) technique over the documents. This technique has been reported as suitable for this task [13]. All parameters were selected to guarantee the best performance of the algorithm. As a non-content baseline, we use the results reported by [11].

**Table 1.** Results over the SemEval13 WSI dataset.

| Algorithm | Parameter | F1 | JI | RI | ARI |
|---|---|---|---|---|---|
| MOO(SCP,PR) | 5 | 0.618 | **0.347** | 0.604 | 0.096 |
| | 10 | 0.679 | 0.332 | 0.605 | 0.128 |
| MOO(PMI,PR) | 5 | 0.646 | 0.352 | 0.604 | 0.128 |
| | 10 | 0.668 | 0.339 | **0.628** | 0.118 |
| MOO(SCP) | 5 | 0.613 | 0.334 | 0.569 | 0.095 |
| | 10 | 0.644 | 0.329 | 0.609 | 0.120 |
| MOO(PMI) | 5 | 0.628 | 0.343 | 0.540 | 0.048 |
| | 10 | 0.630 | 0.330 | 0.552 | 0.059 |
| Hyperlink baseline | 5 | 0.609 | 0.210 | 0.605 | 0.079 |
| | 10 | 0.646 | 0.159 | 0.626 | 0.082 |
| Content baseline | LDA-5 | 0.657 | 0.234 | 0.621 | **0.151** |
| | LDA-10 | **0.716** | 0.168 | 0.626 | 0.131 |

## 6 Results and Discussions

We execute our proposed MOO clustering technique on the SemEval2013 dataset [13]. The parameters of the proposed clustering technique are as follows: $T_{min} = 0.001$, $T_{max} = 100$, $\alpha = 0.9$, $HL = 30$, $SL = 50$ and $iter = 15$. They were determined after conducting a thorough sensitivity study. We perform experiments in four different ways. In the first version, we consider total four objectives: i) SCP based compactness, ii) SCP based separability, iii) hyperlink or pagerank (PR) based compactness and iv) PR based separability. For assigning points to different clusters and also to calculate similarity values during objective function calculation, SCP matrix is used. In the second version of our experiments, we use four objective functions: i) PMI based compactness, ii) PMI based separability, iii) PR based compactness and iv) PR based separability. In this version PMI based similarity measure is used for computing the membership matrix and objective functions. In the third version, we use two objective functions: i) SCP based compactness and ii) SCP based separability. In the fourth version we use two objective functions: i) PMI based compactness ii) PMI based separability.

Results are reported in Table 1. In the table second column (parameter) represents the number of clusters in corresponding version of our experiments. From the results it is evident that for all the validity metrices the first version performs better compared to the third version. It implies that inclusion of hyperlink information makes the clustering algorithm more efficient. Similarly, second version performs better than the fourth version in all aspects. Results also show that both the first two versions of the proposed algorithm (using $SCP$ and $PMI$ based similarity measures, respectively) perform better compared to the approach reported in [11], where only hyperlink information was used. However, the similar situation was not observed when the results of the proposed approach are compared with the content-based baseline. It is important to note that LDA is a strong baseline, and our algorithm shows slight under-performance for F1 (5%) and RI (0.01%). It is more significant for ARI (15%) and on the other hand, MOO outperforms LDA by 30% in terms of JI. Clearly, the use of content has helped in the sense identification, but fails to contribute to their maximum as it is obtained by the use of LDA. Moreover as mentioned in [3], selecting appropriate combination of parameters is very important for the good performance of a particular MOO based approach. Thus a proper sensitivity study is required to conduct to choose the correct values of parameters. In the current approach the same set of parameters as used in [1] is used. But as the targeted task is more complex compared to [1], it will be more interesting to conduct the sensibility analysis further.

## 7 Conclusion

In this paper, we have formulated the problem of WSI within the framework of MOO that combines atypical mixed sources of information. Our proposed approach differs from related work as clustering is performed over multiple objective functions that take into account document content and hyperlink connections. As far as we know, this is the first attempt towards this research direction in WSI studies.

In particular, we proposed the use of similarity metrics based on the frequencies of words in documents ($SCP$ and $PMI$) to evaluate the content similarity and the use of a Jensen-Shannon kernel function based on PageRank to compute the Web pages interconnectivity. Four cluster indices are proposed to guide the optimization process. Results show that the combination of these two different sources outperforms clustering techniques that relay on just one.

## References

1. Acharya, S., Saha, S., Moreno, J.G., Dias, G.: Multi-objective search results clustering. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. pp. 99–108 (2014)
2. Avrachenkov, K., Dobrynin, V., Nemirovsky, D., Pham, S., Smirnova, E.: 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)
3. Bandyopadhyay, S., Saha, S., Maulik, U., Deb, K.: A simulated annealing-based multiobjective optimization algorithm: Amosa. In: IEEE transactions on evolutionary computation. pp. 269–283 (2008)
4. Carpineto, C., Osinski, S., Romano, G., Weiss, D.: A survey of web clustering engines. ACM Computer Survey 41(3), 1–38 (2009)
5. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. Wiley (2009)
6. Di Marco, A., Navigli, R.: Clustering and diversifying web search results with graph-based word sense induction. Computational Linguistics 39(4), 1–43 (2013)
7. Gale, W., Church, K., Yarowsky, D.: One sense per discourse. In: Proceedings of the Workshop on Speech and Natural Language (HLT). pp. 233–237 (1992)
8. Lau, J.H., Cook, P., Baldwin, T.: unimelb: Topic modelling-based word sense induction for web snippet clustering. In: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013) (June 2013)
9. Martins, A., Smith, N., Xing, E., Aguiar, P., Figueiredo, M.: Nonextensive information theoretic kernels on measures. The Journal of Machine Learning Research 10, 935–975 (2009)
10. Meusel, R., Vigna, S., Lehmberg, O., Bizer, C.: Graph structure in the web - revisited. In: Proceedings of the International World Wide Web Conference (WWW). pp. 1–8 (2014)
11. Moreno, J.G., Dias, G.: Pagerank-based word sense induction within web search results clustering. In: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 465–466. JCDL '14 (2014)
12. Morik, K., Kaspari, A., Wurst, M., Skirzynsk, M.: Multi-objective frequent termset clustering. Knowledge Information Systems 30(3), 715–738 (2012)
13. Navigli, R., Vannella, D.: 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)