

Word Sense Induction in the Arabic Language: A Self-Term Expansion Based Approach

David Pinto^(1,2), Paolo Rosso¹, Yassine Benajiba¹,
Anas Ahachad³, Héctor Jiménez-Salazar⁴

¹Natural Language Engineering Lab., RFIA Group
Polytechnic University of Valencia, Spain

²Faculty of Computer Science,
B. Autonomous University of Puebla, Mexico

³Abdelmalek Essaadi University
Tangier, Morocco

⁴Metropolitan Autonomous University
Mexico, DF

{dpinto, proso, ybenajiba}@dsic.upv.es,
{anas.ahachad,hgimenezs}@gmail.com

Abstract. The aim of the word sense induction/discrimination task of natural language processing is to discover the sense associated to each instance of a given ambiguous word. In this paper we present an approach based on clustering of a self-expanded version of the original dataset in order to tackle this particular problem. The self-expansion technique substitutes every term of the original corpus with a set of co-related terms which is calculated by means of pointwise mutual information. Our proposal which was tested for the English language shows a good performance for the Arabic language too, highlighting its language-independent characteristic.

Keywords: Word sense induction, term expansion, clustering.

1 Introduction

Word Sense Induction (WSI) is a particular task of computational linguistics which consists in automatically discover the correct sense for each instance of a given ambiguous word [1]. This problem is closely-related to Word Sense Disambiguation (WSD) [2], however, whereas in WSD the aim is to tag each ambiguous word in a text with one of the senses known a priori, in WSI the aim is to induce the different senses of that word. Typically, the major systems for WSD tackle this task by using two different approaches: corpus-based and knowledge-based. The accuracy of the corpus-based algorithms for WSD is usually proportional to the amount of hand-tagged data available, but the construction of that kind of training data is often difficult for real applications. WSI overcomes this drawback by using clustering algorithms which do not need training data in order to determine the possible sense for a given ambiguous word. The knowledge-based approach uses the ambiguous word context and the information extracted from

ontologies (such as WordNet) in order to disambiguate the different senses of a word, for instance, in [3] a knowledge-based approach which uses the conceptual density technique is presented. Since this viewpoint, WSI is more similar to the knowledge-based WSD than it is to the corpus-based WSD approach.

In this paper we deal with the problem of WSI in the Arabic language. We use a simple technique based on a self-term expansion, which basically constructs a set of co-occurrence terms and, thereafter, it uses this set to expand the target corpus. This approach has been tested in other datasets, with documents written in English, obtaining promising results [4]. The aim of this research work is to analyse the behaviour of the mentioned technique in a completely different language, such as Arabic.

The rest of this paper is structured as follows. The next section introduces the concept of self-term expansion, providing examples in order to easily understand how this technique works. In Section 3 we describe each component involved in the WSI system we developed. Section 4 presents the corpus used in the experiments. The obtained results are shown in Section 5. Finally, the discussion of findings and further work are presented.

2 The Self-Term Expansion Technique

The self-term expansion method consists in replacing terms of a document with a set of co-related terms.

Formally, given a corpus of n documents: $C = \{D_1, D_2, \dots, D_n\}$ with vocabulary \mathcal{V} , a document $D_k \in C$: $D_k = \{W_1, W_2, \dots, W_{|D_k|}\}$, and a set of terms (or words) Co-Related (\mathcal{CR}) to each vocabulary word of C obtained by using the same target dataset: $\mathcal{CR} = \{W_i \overset{\circ}{=} W_j | W_i, W_j \in \mathcal{V}\}^1$, the complete self-term expanded version of C (C') is obtained by replacing each term of the corpus by its co-related terms, that is: $C' = \{D'_1, D'_2, \dots, D'_n\}$ with $D'_k = \{W'_1, W'_2, \dots, W'_{|D_k|}\}$, where $W'_j = \{W_j \cup W_i | W_i \overset{\circ}{=} W_j\}$.

Although the self-term expansion process may be carried out by mean of different ways, often just by using a knowledge database, we particularly consider important to use first the intrinsic information of the target dataset before using external resources.

2.1 Related Work

The expansion of short sentences is not new. In information retrieval, for instance, the expansion of query terms is a very investigated topic which has shown to improve results with respect to when query expansion is not employed [5,6,7,8,9].

The availability of Machine Readable Resources (MRR) like “Dictionaries”, “Thesauri” and “Lexicons” has allowed to apply term expansion to other fields of natural language processing like WSD. In [10] we may see the typical example

¹ We will use the symbol $\overset{\circ}{=}$ to mathematically represent the co-relation operator.

of using an external knowledge database for determining the correct sense of a word given in some context. In this approach, every word close to the one we would like to determine its correct sense is expanded with its different senses by using the WordNet ontology [11]. Then, an overlapping factor is calculated in order to determine the correct sense of the ambiguous word. Different other approaches have made use of a similar procedure. By using dictionaries, the proposals presented in [12,13,14] are the most successful in WSD. Yarowsky [15] used instead thesauri for his experiments. Finally, in [16,17,10] the use of lexicons in WSD has been investigated. Although in some cases the knowledge resource seems not to be used strictly for term expansion, the application of co-occurrence terms is included in their algorithms.

Like in information retrieval, in WSD the application of term expansion with co-related terms has shown to improve the baseline results if we carefully select the external resource to use, with a priori knowledge of the domain. Evenmore, we have to be sure that the Lexical Data Base (LDB) has been suitably constructed. Due to the last facts, we consider that the use of a self automatically constructed LDB (using the same test corpora), may be of high benefit. This assumption is based on the intrinsic properties extracted from the corpus itself. Our proposal is somehow related with the approaches presented in [18] and [19], where words are also expanded with co-occurrence terms for word sense discrimination. The main difference consists in the use of the same corpus for constructing the co-occurrence list and not of an external resource.

2.2 Construction of the Co-Occurrence List

In literature, co-occurrence terms is the most common technique used for the automatic construction of LDBs [8,20]. On the one hand, a simple approach may use n -grams, which allow us to predict a word from previous words in a sample of text. The frequency of each n -gram is calculated and then filtered according to some threshold. The resulting n -grams constitute a LDB which may be used as an “expansion dictionary” for each term. On the other hand, an information theory-based co-occurrence measure is discussed in [21]. This measure is named pointwise Mutual Information (MI), and its applications for finding collocations are analysed by determining the co-occurrence degree among two terms. This may be done by calculating the ratio between the number of times that both terms appear together (in the same context and not necessarily in the same order) and the product of the number of times that each term occurs alone. Given two terms x_1 and x_2 , the pointwise mutual information between x_1 and x_2 can be calculated as follows:

$$MI(x_1, x_2) = \log_2 \frac{P(x_1 x_2)}{P(x_1) \times P(x_2)}$$

For instance, if we know the words “Computer” and “Science” occurs, respectively, with frequency 70 and 60, and the term “Computer Science” occurs with frequency 20 in a corpus of 15,000,000 tokens, then we may calculate the

pointwise mutual information among these two words as:

$$MI(\text{Computer, Science}) = \log_2 \frac{\frac{20}{15,000,000}}{\frac{70}{15,000,000} \times \frac{60}{15,000,000}} \approx 16.12$$

The numerator could be modified in order to take into account only bigrams, as presented in [22], where an improvement of clustering short texts in narrow domains (i.e. domains with a high degree of overlapping between their vocabularies) has been obtained. We determined that the single occurrence of each term should be at least three (see [21]), whereas the maximum separation among the two terms was five.

We have used the pointwise MI for obtaining a co-occurrence list from the same target dataset. This list is then used to expand every term of the original corpus. Since the co-occurrence formula captures relations between related terms, it is possible to see that the self-term expansion magnifies the noisy in a lower degree than it does for the meaningful information. Therefore, the execution of the clustering algorithm in the expanded corpus should outperform the one executed over the non-expanded data.

In order to fully appreciate the self-term expansion method, in Table 1 we show the co-occurrence list for some words related with the verb “kill” calculated from the English language corpus used in the “Evaluating Word Sense Induction and Discrimination Systems” task of the SemEval 2007 workshop [1,4]. Since the MI is calculated after preprocessing the corpus, we present the stemmed version of the terms.

Table 1. An example of co-occurrence terms

Word	Co-occurrence terms
soldier	kill
rape	women think shoot peopl old man kill death beat
grenad	today live guerrilla fight explod kill
death	shoot run rape person peopl outsid murder life lebanon kill convict...
temblor	tuesdai peopl least kill earthquak

3 The Word Sense Induction System

Given a set of ambiguous words, each one with a set of instances, the goal of the WSI system is to discriminate among all the instances and automatically discover the sense each instance belongs to [4]. In Figure 1 we may see a diagram of the described WSI process. The ambiguous word “construct” and a set of correspondant instances feed the WSI system, which outputs a set of discovered

senses. In this figure we have used the same format provided in the “Evaluating Word Sense Induction and Discrimination Systems” task of the SemEval 2007 workshop². The ambiguous word are enclosed by two tags: `<lexelt>` and `</lexelt>`. Each instance of a given ambiguous word contains a unique identifier (*id*) and it is enclosed by the tags `<head>` and `</head>` in the given paragraph.

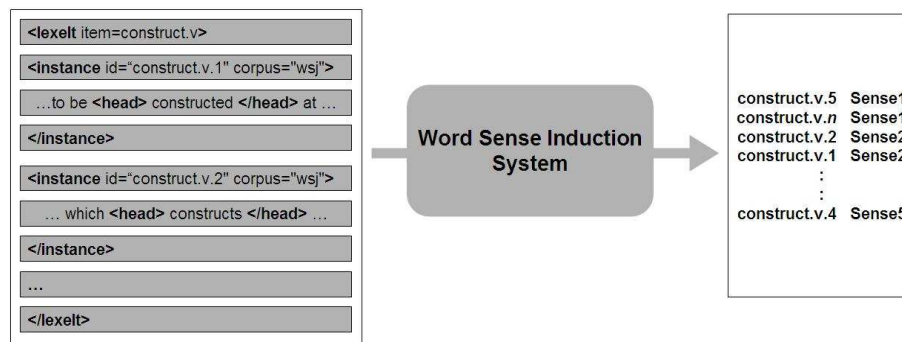


Fig. 1. The word sense induction task

The developed WSI system is composed of different modules which are presented in Figure 2. Two are the basic components which are involved: the self-term expansion technique and the clustering method. The former contains two basic submodules: the co-occurrence list constructor which uses pointwise mutual information, and the submodule which expands the terms of the input data. The latter module employed calculates a similarity matrix over the expanded corpus for the clustering method.

We selected the unsupervised KStar clustering method [23] for the experiments, defining the average of similarities among all the sentences for a given ambiguous word as the stop criterion in the clustering process. The input similarity matrix for the clustering method was calculated by using the Jaccard coefficient, which may be expressed as follows. Given two documents D_1 and D_2 , the Jaccard coefficient (Jaccard) among them is:

$$\text{Jaccard}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

The Jaccard coefficient will get a normalized value between zero and one. A value of one will be obtained when the two given documents contain the same set of terms (words), whereas a value of zero will be obtained when no identical words at all are shared by the given documents.

² <http://nlp.cs.swarthmore.edu/semeval/tasks/task02/description.shtml>

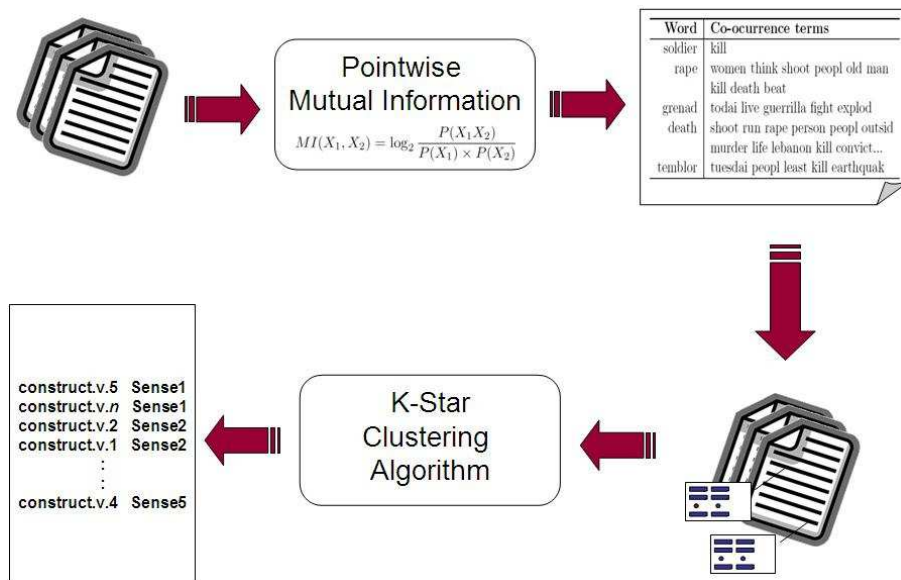


Fig. 2. The main components of the proposed WSI system

4 Dataset

For the experiments carried out in this research work, we have used the dataset prepared for the Arabic tasks of the SemEval workshop³. A set of 509 ambiguous words (379 nouns and 130 verbs) were provided. We preprocessed this original dataset by eliminating punctuation symbols and Arabic stopwords. The experiments were carried out by using a tokenized (with segmentation) version of the target corpus. The complete characteristics of the used corpus are described in Table 2.

Table 2. Characteristics of the Arabic corpus used in the WSI experiments

Characteristic	Value
Size	343 Kbytes
Ambiguous words	509
Nouns	379
Verbs	130
Instances (in average)	1025

³ <http://nlp.cs.swarthmore.edu/semeval/tasks/task18/description.shtml>

5 Experimental Results

In order to determine the efficacy of the approach presented below we have performed our preliminary experiment on the Semeval Arabic task corpus. This corpus is already tokenized, i.e., we can find the root morpheme and the affixes of each word separated. However, it is impossible to compute the precision and recall of our approach because the gold standard file is not released yet. Therefore, in this section we will present an evaluation based on our own judgment of what we have obtained.

The examples showed in this section are given first in Buckwalter transliterated characters and, thereafter, in Arabic language. The Buckwalter transliteration was developed by Tim Buckwalter for practical storage, display and email transmission of Arabic text in environments where the display of genuine Arabic characters is not possible or convenient [24].

Some examples show that the approach performed quite well on Arabic data. For instance, the Arabic word “كل” (“kl” in the Buckwalter transliterated characters) in the data has two different meanings: the first one is “all” and the second is “every”. The obtained results of the performed discrimination for this particular ambiguous word are shown as follows.

Sense “all (kl)”:

- 1) w >wDHt An " AltEAwn AlHAly byn Albldyn ysy b xTY
Hvyvp w fy kl AlmjAlAt " .
- 2) w qAl AlsA}H AlbryTAny jwrj dwd , m\$yrAF ALY Alfndq
AlAnyq AlmTl ELY AlbHr : " nElm >n hm yqymwn fy h*A
Alfndq w lA ymknn ny Alqwl <n nA n\$Er b {rtyAH mE wjwd
hm w kl h*A AlEdd mn rjAl Al\$rTp Hwl nA " .
- 3) w qAl AlsA}H AlAlmAny bwl hwfmAn , mtHdvAF En AlHrAs
Almtmrkzyn ELY sTH Alfndq : " AEtqd An mn Algryb AHATp
AlmkAn b kl AjrA'At AlAmn h*h .
- 4) HtY lA ysa' tfsyr EbArp " rfD AlEnf b kl >\$kAl
h " AlwArdp fy byAn ...

Sense “all (كل)”:

واوضحت ان التعاون الحالي بين البلدين يسير بخطى حثيثة وفي كل المجالات
وقال السائح البريطاني جورج داود مشيراً الى الفندق الانيق المطل على البحر
نعلم انهم يقيمون في هذا الفندق ولا يمكنني القول اننا نشعر بارتياح مع وجودهم
وكل هذا العدد من رجال الشرطة حولنا
وقال السائح الاناني بول هوفمان متحدثاً عن الخراس المتمركزين على سطح الفندق
اعتقد ان من الغريب احاطة المكان بكل اجراءات الامن هذه
حتى لايساء تفسير عبارة رفض العنق بكل اشكاله الواردة في بيان

Sense “every (kl)”:

- 1) -LRB- . . . -RRB- f Alkl hnA tHt AlmrAqbp h*h AlAyAm " .

2) w zAr Aldwry klA mn AlArdn w lbnAn w swryA fy ATAr
 jwlp tsthdH H\$d AlmEARdp AlErbypl >y Hmlp qd t\$n hA
 AlwLAyAt AlmtHdp ELY AlErAq .

Sense “every (كل)”:

قالكل هنا تحت المراقبة هذه الايام
 وزار الدوري كلامن الاردن ولبنان وسورية في اطار جولة تستهدف حشد المعارضة
 العربية لاي حملة قد تشنها الولايات المتحدة الامركية

Meanwhile, sometimes our method tends to discriminate several senses of a word even if all the instances of the word express the same sense. In the following example, for instance, all the samples of the Arabic word “جندى” (soldier - jndy) have the same sense. However, our method discriminates the first instance as having a different sense mainly because it appears in a quite different context.

Word “soldier (jndy)”:

1) w >fAd ms&wln hnwd An jndyyn lqyA Htf hmA w >n
 vlAvp |xryn jrHwA xlAl ATlAq nAr fy mstwdE l *xyrp Aljy\$
 fy wLAyp jAmw w k\$myr Alhndyp .

2) w qAl DAbT \$rTp An vwArA dhmwA AlmstwdE fy mqATEp
 bwn\$ ELY msAfp 250 kylwtrA \$mAl jAmw AlEASmp Al\$twyp l
 AlwLAyp w >lqWA AlnAr f qtl jndyAn fwrAF w >Syb vlAvp
 qbl frAr Almqtlyn .

3) w nfY ms&wl |xr fy AlHkwmp HSwl hjwm l Almt\$ddyn w
 qAl An jndyAF hw Al*y >lq AlnAr ELY zmlA} h .

Word “soldier (جندى)”:

واقفاد مسؤولون هنود ان جنديين لقيتا حتفهما وان ثلاثة اخرين جرحوا خلال اطلاق
 نار في مستودع لدخيرة الجيش في ولاية جامو وكشمير الهندية
 وقال سابط شرطة ان ثوارا دهموا المستودع في مقاطعة بونش على مسافة 250 كيلومترا
 شمال جامو العاصمة الشتوية للولاية والقوا النار فقتل جنديان فورا واصيب ثلاثة
 قتل فرار المقاتلين
 ونفى مسؤول اخر في الحكومة حصول هجوم للمتشددين وقال ان جنديا هو الذي القى
 النار على زملائه

After examining our results we have observed that the more instances we have for a given ambiguous word (and thus more contextual information), the better our method discriminates the different senses of the given instances. In Figure 3 we may see the example of the noun “President” used in four different sentences. In the first sentence the word president is used to express the “Prime Minister” which is said in Arabic “Ministers President”, whereas in the other three sentences “president” expresses “head of nation or country”. In this case the approach that we have used succeeded in discriminating the two mentioned senses.

Meanwhile, when we have poor context information (few senses using the ambiguous word) our method is not able to discriminate the different senses of

و قال زعيم آخر في المؤتمر هو عمر فاروق , ان الكرة الان في الملعب الهندي . و ان على **رئيس** الوزراء الهندي أتال بيهاري فاجاي الان ان يرد على مبادرة
And another leader of the conference was Umar Farooq, he said that the ball is now in India's stadium and that the Indian **Prime Minister** Atal Bihari Vajpayee now had "to respond to the initiative

و رحبت " جبهة تحرير جامو و كشمير " الانفصالية ب خطاب **الرئيس** الباكستاني و خصوصا التأكيد مجددا ل الدعم السياسي و المعنوي و الدبلوماسي لالكشميريين
And Liberation Front "the Jammu and Kashmir separatist" welcomed the Pakistani **President** speech and especially the reiteration to support for the political, moral and diplomatic the Koshemurien

واصلت نيودلهي الضغط على اسلام اباد لقمع الجماعات الاسلامية المتشددة , فأعلن وزير الدفاع الهندي جورج فرناندز ان القوات الهندية المحتشدة على الحدود مع باكستان لن تسحب اذا لم يترجم **الرئيس** الباكستاني برويز مشرف تعهداته كبح جماح الاسلاميين الى افعال
New Delhi continues to put pressure on Islamabad to suppress extremist Islamic groups, and the Indian Defense Minister George Fernandes announced that Indian forces mobilized on the borders with Pakistan would only withdraw if the Pakistani **President** Pervez Musharraf will curb the Islamists as he promised.

و أكد نائب **الرئيس** العراقي طه ياسين رمضان الاحد الماضي ان بلاده لن تسمح بعودة مفتشي الاسلحة
And Iraq's Vice **President** Taha Yassin Ramadan said last Sunday that the country will not allow the return of weapons inspectors

Fig. 3. Samples of the noun "President"

the word and tends to cluster them in the same group. For instance, the use of the verb "to see" in the first sentence of the example given in Figure 4 expresses the opinion of somebody about something, and in the second one the verb "to see" is used to express what a man sees with his eyes. However, there are very few sentences which contain the verb "to see" in the corpus we have used and, therefore, our system did not obtain enough information about the contexts in which this verb would possibly appear.

اسير على الشاطئ يوميا و عندما ارى الجنود على السطح اشعر بنوع من عدم الارتياح ... اعتقد أنه كان ممكنا اختيار مكان آخر لهم
I walk on the beach every day and when I see the soldiers on the roof I feel a kind of uncomfortable ... I think it was possible to choose another place for them

و أضاف : بعد المناقشات , رأأت اللجنة أن لا بد من تنفيذ بنود مبادرة السلام العربية المنبثقة من القمع , و التي تتبنى الثوابت في مواجهة المؤامرة الصهيونية ضد الشعب الفلسطيني
And added : Following the discussions, the Committee sees that we must implement the provisions of the Arab peace initiative which were born from repression, and which adopts the constants in confronting the Zionist plot against the Palestinian people.

Fig. 4. Samples of the verb "to see"

6 Conclusions and Further Work

We have presented a language-independent system for word sense induction/discrimination. The approach is based on self-term expansion, and it uses the point-wise mutual information for calculating the set of co-related terms needed in the term expansion process of the original corpus. Our method uses the KStar clustering method in order to induce all the possible senses for each ambiguous word of the target dataset. The preliminary experiments show a good performance in Arabic as well as it was with the English language.

The tokenization performed on the Semeval corpus by the task organizers was only a partial tokenization as they have kept the Arabic definite article “Al” joint to the words. This partial tokenization might be positive for other NLP tasks, however we consider that the method presented in this paper would perform better if we change Mona Diab’s tokenizer in order to take into consideration the definite article. We are also in contact with SemEval 2007 workshop organizers in order to perform further experiments as soon as they will be able to provide more corpora.

Acknowledgements

This work has been partially supported by the PCI-AECI A/7067/06 and MCyT TIN2006-15265-C06-04 projects, as well as by the BUAP-701 PROMEP/103.5/-05/1536 grant and the “Programa de Apoyo a la Investigación y Desarrollo” (PAID-06-06) of the Polytechnic University of Valencia.

References

1. Agirre, E., Soroa, A.: Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In: Proceedings of the SemEval Workshop, Prague, Czech Republic, The Association for Computational Linguistics (2007) 7–12
2. Ide, N., Véronis, J.: Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics* **24**(1) (1998) 2–40
3. Buscaldi, D., Rosso, P., Masulli, F.: The upv-unige-ciaosenso wsd system. In: Proceedings of the Senseval-3 Workshop, Barcelona, Spain, The Association for Computational Linguistics (2004) 77–82
4. Pinto, D., Rosso, P., Jiménez-Salazar, H.: Upv-si: Word sense induction using self term expansion. In: Proceedings of the SemEval Workshop, Prague, Czech Republic, The Association for Computational Linguistics (2007) 430–433
5. Qiu, Y., Frei, H.P.: Concept based Query Expansion. In: ACM SIGIR on R&D in information retrieval, ACM Press (1993) 160–169
6. Ruge, G.: Experiments on linguistically-based term associations. *Information Processing & Management* **28**(3) (1992) 317–332
7. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. New York: ACM Press; Addison-Wesley (1999)
8. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic (1994)

9. Rijsbergen, C.J.V.: Information Retrieval, 2nd edition. Dept. of Computer Science, University of Glasgow, Scotland (1979)
10. Banerjee, S., Pedersen, T.: An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In: CICLing 2002 Conference. Volume 3878 of LNCS., Springer-Verlag (2002) 136–145
11. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
12. Lesk, M.: Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In: ACM SIGDOC Conference, ACM Press (1986) 24–26
13. Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T., Slator, B.: Providing machine tractable dictionary tools. *Machine Translation* **5**(2) (1990) 99–154
14. Nancy, I., Véronis, J.: Mapping dictionaries: A spreading activation approach. In: 6th Annual Conference of the Centre for the New Oxford English Dictionary. (1990) 52–64
15. Yarowsky, D.: Word-sense disambiguation using statistical models of Rogets categories trained on large corpora. In: 14th Conference on Computational Linguistics, The Association for Computational Linguistics (1992) 454–460
16. Sussna, M.: Word sense disambiguation for free-text indexing using a massive semantic network. In: 2nd International Conference on Information and Knowledge Management. (1993) 67–74
17. Resnik, P.: Disambiguating Noun Groupings with Respect to WordNet Senses. In: 3rd Workshop on Very Large Corpora, The Association for Computational Linguistics (1995) 54–68
18. Schütze, H.: Automatic word sense discrimination. *Computational Linguistics* **24**(1) (1998) 97–123
19. Purandare, A., Pedersen, T.: Word sense discrimination by clustering contexts in vector and similarity spaces. In: Proceedings of the Conference on Computational Natural Language Learning, Boston, MA (2004) 41–48
20. Frakes, W.B., Baeza-Yates, R.A.: Information Retrieval: Data Structures & Algorithms. Prentice-Hall (1992)
21. Manning, D.C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (2003) Revised version, May 1999.
22. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering abstracts of scientific texts using the transition point technique. In: CICLing. Volume 3878 of LNCS., Springer-Verlag (2006) 536–546
23. Shin, K., Han, S.Y.: Fast clustering algorithm for information organization. In: CICLing. Volume 2588 of LNCS., Springer-Verlag (2003) 619–622
24. Buckwalter, T.: Issues in arabic orthography and morphology analysis. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, Italy (2004)