Kyoto: An Integrated System for Specific Domain WSD

Aitor Soroa, Eneko Agirre, Oier Lopez de Lacalle
University of the Basque Country
a.soroa@ehu.es

Monica Monachini
Istituto di Linguistica Computazionale
monica.monachini@ilc.cnr.it

Jessie Lo, Shu-Kai Hsieh
National Taiwan Normal University
shukai@ntnu.edu.tw

Wauter Bosma, Piek Vossen
Vrije Universiteit
{p.vossen,w.bosma}@let.vu.nl

Abstract

This document describes the preliminary release of the integrated Kyoto system for specific domain WSD. The system uses concept miners (Tybots) to extract domain-related terms and produces a domain-related thesaurus, followed by knowledge-based WSD based on wordnet graphs (UKB). The resulting system can be applied to any language with a lexical knowledge base, and is based on publicly available software and resources. Our participation in Semeval task #17 focused on producing running systems for all languages in the task, and we attained good results in all except Chinese. Due to the pressure of the time-constraints in the competition, the system is still under development, and we expect results to improve in the near future.

1 Introduction

In this paper we describe the participation of the integrated Kyoto system on the “SemEval-2010 task #17: All-words Word Sense Disambiguation on a Specific Domain” task (Agirre et al., 2010). The goal of our participation was to evaluate the preliminary release of the integrated system for specific domain WSD developed for the Kyoto project[1]. Besides, we wanted to test the performance of our domain specific WSD system (Agirre et al., 2009) on this test set, and to integrate the thesaurus construction software (Tybots) developed for the project. The system can be run for any language and domain if provided with a lexical knowledge base and some background documents on the domain.

We will first present the components of our system, followed by the experimental design and the results. Finally, the conclusions are presented.

2 The Kyoto System for Domain Specific WSD

We will present in turn UKB, the Tybots, and the lexical knowledge-bases used.

2.1 UKB

UKB is a knowledge-based unsupervised WSD system which exploits the structure of an underlying Language Knowledge Base (LKB) and finds the most relevant concepts given an input context (Agirre and Soroa, 2009). UKB starts by taking the LKB as a graph of concepts $G = (V, E)$ with a set of vertices V derived from LKB concepts and a set of edges E representing relations among them. Giving an input context, UKB applies the so called *Personalized PageRank* (Haveliwala, 2002) over it to obtain the most representative senses for the context.

PageRank (Brin and Page, 1998) is a method for scoring the vertices V of a graph according to each node’s structural importance. The algorithm can be viewed as random walk process that postulate the existence of a particle that randomly traverses the graph, but at any time may jump to a new vertex with a given *damping factor* (also called *teleport probability*). After PageRank calculation, the final weight of node i represents the proportion of time that a random particle spends visiting node i after a sufficiently long time. In standard PageRank, the teleport vector is chosen uniformly, whereas for Personalized PageRank it is chosen from a nonuniform distribution of nodes, specified by a *teleport vector*.

UKB concentrates the initial probability mass of the teleport vector in the words occurring in the context of the target word, causing all random jumps on the walk to return to these words and thus assigning a higher rank to the senses linked to these words. Moreover, the high rank of the words

[1] http://www.kyoto-project.eu

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 417–420,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

spreads through the links in the graph and make all the nodes in its vicinity also receive high ranks. Given a target word, the system checks which is the relative ranking of its senses, and the WSD system would output the one ranking highest.

UKB is very flexible and can be use to perform WSD on different settings, depending on the context used for disambiguating a word instance. In this paper we use it to perform general and domain specific WSD, as shown in section 3. PageRank is calculated by applying an iterative algorithm until convergence below a given threshold is achieved. Following usual practice, we used a damping value of 0.85 and set the threshold value at 0.001. We did not optimize these parameters.

2.2 Tybots

Tybots (Term Yielding Robots) are text mining software that mine domain terms from corpus (e.g. web pages), organizing them in a hierarchical structure, connecting them to wordnets and ontologies to create a semantic model for the domain (Bosma and Vossen, 2010). The software is freely available using Subversion [2]. Tybots try to establish a view on the terminology of the domain which is as complete as possible, discovering relations between terms and ranking terms by domain relevance.

Preceding term extraction, we perform tokenization, part-of-speech tagging and lemmatization, which is stored in Kyoto Annotation Format (KAF) (Bosma et al., 2009). Tybots work through KAF documents, acquire domain relevant terms based on the syntactic features, gather cooccurrence statistics to decide which terms are significant in the domain and produce a thesaurus with sets of related words. Section 3.3 describes the specific settings that we used.

2.3 Lexical Knowledge bases

We used the following wordnets, as suggested by the organizers:

WN30g: English WordNet 3.0 with gloss relations (Fellbaum, 1998).

Dutch: The Dutch LKB is part of the Cornetto database version 1.3 (Vossen et al., 2008). The Cornetto database can be obtained from the Dutch/Flanders Taalunie[3]. Cornetto comprises taxonomic relations and equivalence relations from both WordNet 2.0 and 3.0. Cornetto concepts are mapped to English WordNet 3.0.

	#entries	#synsets	#rels.	#WN30g
		Monolingual		
Chinese	8,186	14,243	20,433	20,584
Dutch	83,812	70,024	224,493	83,669
Italian	46,724	49,513	65,567	52,524
WN30g	147,306	117,522	525,351	n/a
		Bilingual		
Chinese-eng	8,186	141,561	566,368	
Dutch-eng	83,812	188,511	833,513	
Italian-eng	46,724	167,094	643,442	

Table 1: Wordnets and their sizes (entries, synsets, relations and links to WN30g).

Italian: Italwordnet (Roventini et al., 2003) was created in the framework of the EuroWordNet, employs the same set of semantic relations used in EuroWordNet, and includes links to WordNet 3.0 synsets.

Chinese: The Chinese WordNet (Version 1.6) is now partially open to the public[4] (Tsai et al., 2001). The Chinese WordNet is also mapped to WordNet 3.0.

Table 1 shows the sizes of the graphs created using each LKB as a source. The upper part shows the number of lexical entries, synsets and relations of each LKB. It also depicts the number of links to English WordNet 3.0 synsets.

In addition, we also created bilingual graphs for Dutch, Italian and Chinese, comprising the original monolingual LKB, the links to WordNet 3.0 and WordNet 3.0 itself. We expected this richer graphs to perform better performance. The sizes of the bilingual graphs are shown in the lower side of Table 1.

3 Experimental setting

All test documents were lemmatized and PoS-tagged using the linguistic processors available within the Kyoto project. In this section we describe the submitted runs.

3.1 UKB parameters

We use UKB with the default parameters. In particular, we don't use dictionary weights, which in the case of English come from annotated corpora. This is done in order to make the system fully unsupervised. It's also worth mentioning that in the default setting parts of speech were not used.

[2] http://kyoto.let.vu.nl/svn/kyoto/trunk

[3] http://www.inl.nl/nl/lexica/780

[4] http://cwn.ling.sinica.edu.tw

RANK	RUN	P	R	R-NOUN	R-VERB
Chinese					
-	*1sense*	0.562	0.562	0.589	0.518
1	*Best*	0.559	0.559	-	-
-	*Random*	0.321	0.321	0.326	0.312
4	kyoto-3	0.322	0.296	0.257	0.360
3	kyoto-2	0.342	0.285	0.251	0.342
5	kyoto-1	0.310	0.258	0.256	0.261
Dutch					
1	kyoto-3	0.526	0.526	0.575	0.450
2	kyoto-2	0.519	0.519	0.561	0.454
-	*1sense*	0.480	0.480	0.600	0.291
3	kyoto-1	0.465	0.465	0.505	0.403
-	*Random*	0.328	0.328	0.350	0.293
English					
1	*Best*	0.570	0.555	-	-
-	*1sense*	0.505	0.505	0.519	0.454
10	kyoto-2	0.481	0.481	0.487	0.462
22	kyoto-1	0.384	0.384	0.382	0.391
-	*Random*	0.232	0.232	0.253	0.172
Italian					
1	kyoto-3	0.529	0.529	0.530	0.528
2	kyoto-2	0.521	0.521	0.522	0.519
3	kyoto-1	0.496	0.496	0.507	0.468
-	*1sense*	0.462	0.462	0.472	0.437
-	*Random*	0.294	0.294	0.308	0.257

Table 2: Overall results of our runs, including precision (P) and recall (R), overall and for each PoS. We include the First Sense (1sense) and random baselines, as well as the best run, as provided by the organizers.

3.2 Run1: UKB using context

The first run is an application of the UKB tool in the standard setting, as described in (Agirre and Soroa, 2009). Given the input text, we split it in sentences, and we disambiguate each sentence at a time. We extract the lemmas which have an entry in the LKB and then apply Personalized PageRank over all of them, obtaining a score for every concept of the LKB. To disambiguate the words in the sentence we just choose its associated concept (sense) with maximum score.

In our experiments we build a context of at least 20 content words for each sentence to be disambiguated, taking the sentences immediately before when necessary. UKB allows two main methods of disambiguation, namely *ppr* and *ppr_w2w*. We used the latter method, as it has been shown to perform best.

In this setting we used the monolingual graphs for each language (cf. section 2.3). Note that in this run there is no domain adaptation, it thus serves us as a baseline for assessing the benefits of applying domain adaptation techniques.

3.3 Run2: UKB using related words

Instead of disambiguating words using their context of occurrence, we follow the method described in (Agirre et al., 2009). The idea is to first obtain a list of related words for each of the target words, as collected from a domain corpus. On a second step each target word is disambiguated using the N most related words as context (see below). For instance, in order to disambiguate the word *environment*, we would not take into account the context of occurrence (as in Section 3.2), but we would use the list of most related words in the thesaurus (e.g. "*biodiversity, agriculture, ecosystem, nature, life, climate, . . .*"). Using UKB over these contexts we obtain the most predominant sense for each target word in the domain(McCarthy et al., 2007), which is used to label all occurrences of the target word in the test dataset.

In order to build the thesaurus with the lists of related words, we used Tybots (c.f. section 2.2), one for each corpus of the evaluation dataset, i.e. Chinese, Dutch, English, and Italian. We used the background documents provided by the organizers, which we processed using the linguistic processors of the project to obtain the documents in KAF. We used the Tybots with the following settings. We discarded co-occurring words with frequencies below 10[5]. Distributional similarity was computed using (Lin, 1998). Finally, we used up to 50 related words for each target word.

As in run1, we used the monolingual graphs for the LKBs in each language.

3.4 Run3: UKB using related words and bilingual graphs

The third run is exactly the same as run2, except that we used bilingual graphs instead of monolingual ones for all languages other than English (cf. section 2.3). There is no run3 for English.

4 Results

Table 2 shows the results of our system on the different languages. We will analyze different aspects of the results in turn.

Domain adaptation: Using Personalized Pagerank over related words (run2 and run3) consistently outperforms the standard setting (run1) in all languages. This result is consistent with

[5] In the case of Dutch we did not use any threshold due to the small size of the background corpus.

our previous work on English (Agirre et al., 2009), and shows that domain adaptation works for knowledge-based systems.

Monolingual vs. Bilingual graphs: As expected, we obtained better results using the bilingual graphs (run3) than with monolingual graphs (run2), showing that the English WordNet has a richer set of relations, and that those relations can be successfully ported to other languages. This confirms that aligning different wordnets at the synset level is highly beneficial.

Overall results: the results of our runs are highly satisfactory. In two languages (Dutch and Italian) our best runs perform better than the first sense baseline, which is typically hard to beat for knowledge-based systems. In English, our system performs close but below the first sense baseline, and in Chinese our method performed below the random baseline.

The poor results obtained for Chinese can be due the LKB topology; an analysis over the graph shows that it is formed by a large number of small components, unrelated with each other. This 'flat' structure heavily penalizes the graph based method, which is many times unable to discriminate among the concepts of a word. We are currently inspecting the results, and we don't discard bugs, due to the preliminary status of our software. In particular, we need to re-examine the output of the Tybot for Chinese.

5 Conclusions

This paper describes the results of the preliminary release of he integrated Kyoto system for domain specific WSD. It comprises Tybots to construct a domain-related thesaurus, and UKB for knowledge-based WSD based on wordnet graphs. We applied our system to all languages in the dataset, obtaining good results. In fact, our system can be applied to any language with a lexical knowledge base, and is based on publicly available software and resources. We used the wordnets and background texts provided by the organizers of the task.

Our results show that we were succesful in adapting our system to the domain, as we managed to beat the first sense baseline in two languages. Our results also show that adding the English WordNet to the other language wordnets via the available links is beneficial.

Our participation focused on producing running systems for all languages in the task, and we attained good results in all except Chinese. Due to the pressure and the time-constraints in the competition, the system is still under development. We are currently revising our system for bugs and finetuning it.

Acknowledgments

This work task is partially funded by the European Commission (KYOTO ICT-2007-211423), the Spanish Research Department (KNOW-2 TIN2009-14715-C04-01) and the Basque Government (BERBATEK IE09-262).

References

E. Agirre and A. Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of EACL09*, pages 33–41. Association for Computational Linguistics.

E. Agirre, O. López de Lacalle, and A. Soroa. 2009. Knowledge-based wsd on specific domains: Performing better than generic supervised wsd. In *Proceedigns of IJCAI. pp. 1501-1506."*.

E. Agirre, O. López de Lacalle, C. Fellbaum, S.K. Hsieh, M. Tesconi, M. Monachini, P. Vossen, and R. Segers. 2010. Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Same volume.*

W. E. Bosma and P. Vossen. 2010. Bootstrapping language neutral term extraction. In *Proceedings of LREC2010*, May.

W. E. Bosma, P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi. 2009. KAF: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation.*

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7).

C. Fellbaum. 1998. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.

T. H. Haveliwala. 2002. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on WWW*, pages 517–526, New York, NY, USA. ACM.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL98*, Montreal, Canada.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4).

A. Roventini, A. Alonge, F. Bertagna, N. Calzolari, J. Cancila, C. Girardi, B. Magnini, R. Marinelli, M. Speranza, and A. Zampolli. 2003. Italwordnet: building a large semantic database for the automatic treatment of Italian. *Linguistica Computazionale, Special Issue (XVIII-XIX)*, pages 745–791.

B.S. Tsai, C.R. Huang, S.c. Tseng, J.Y. Lin, K.J. Chen, and Y.S. Chuang. 2001. Definition and tests for lexical semantic relations in Chinese. In *Proceedings CLSW 2001.*

P. Vossen, I. Maks, R. Segers, H. van der Vliet, and H. van Zutphen. 2008. The cornetto database: the architecture and alignment issues. In *Proceedings GWC 2008*, pages 485–506.