Using the Multilingual Central Repository for Graph-Based Word Sense Disambiguation

Eneko Agirre and Aitor Soroa
<a.soroa@ehu.es>

University of the Basque Country

LREC, Marrakesh 2008

Introduction

- WSD: assign a sense to a word in a particular context
- Supervised WSD performs best
  - but needs large amounts of hand-tagged data
- Knowledge-based WSD
  - Exploit information present on a LKB
  - No further corpus evidence

Knowledge-based WSD

- Traditional approach:
  - Assign a sense to an ambiguous word by comparing each of its senses with those of the surrounding context
  - Some semantic similarity metric used for calculating the relatedness among senses
  - Due to combinatorial explosion, words are disambiguated individually
- Graph based methods
  - Graph-based techniques to exploit the structural properties of the graph underlying the LKB
  - Find globally optimal solutions given the relations between entities
  - Disambiguate large portions of text in one go

Main goal of the work

- Novel graph-based method for performing unsupervised WSD
- The method is independent of underlying LKB
  - Applied to Multilingual Central Repository (MCR)
- Evaluate separate and combined performance of several relation types of the MCR

Outline

A graph algorithm for knowledge-based WSD

- Represent the LKB as a graph
  - Nodes are the concepts (v_i)
  - Edges are relations among concepts (e_{ij})
- Given an input context
  - $W_i \; i = 1 \ldots m$: content words (nouns, verbs, adjectives and adverbs)
  - $Synsets_i = \{v_{i1}, \ldots, v_{in}\}$: synsets associated to word i
- Two steps for WSD
  1. Extract a representative subgraph: disambiguation subgraph
  2. Find the “best” synsets of the subgraph

Extracting the disambiguation subgraph

- Subgraph extraction:
  - For each word $W_i,\ i = 1 \ldots m$
  - For each synset $v_{i1} \ldots v_{in}$ input word W_i
    - Find the shortest paths from v_{ij} to synsets of rest of words (BFS search)
  - Create subgraph by joining all minimum distance paths
- The vertices and relations of the subgraph are particularly relevant for a given input context.

Identifying the best synsets: PageRank

- Google's PageRank (Brin and Page, 1998): model a random walk on the graph
  - A walker takes random steps
  - Converges to a stationary distribution of probabilities
- $G = (V, E)$ a graph
  - $In(V_i)$ = nodes pointing to V_i
  - d_j = degree of node v_j

$$\mathrm{PR}(V_i) = (1 - \alpha) + \alpha \sum_{j \in In(V_i)} \frac{1}{d_j} \mathrm{PR}(V_j)$$

Usually $\alpha = 0.85$. Models random jumps.

Identifying the best synsets: PageRank

- PageRank ranks vertices according to their structural importance on the graph
- Apply PageRank over disambiguation subgraph
- Select the synsets with maximum rank for each input word
    - In case of ties, select all synsets with same rank

Outline

Multilingual Central Repository (MCR)

- Knowledge base built whithin the MEANING project
  - Multilingual interface for integrating and distributing all the knowledge acquired in the project
- Current version: $1,500,000$ relations
  - Most of them automatic
- MCR integrates
  - ILI based on WN1.6
  - EWN Base Concepts
  - MultiWordNet Domains (MWND)
  - Local WordNets connected to the ILI
    - English WN1.5, 1.6, 1.7, 1.7.1
    - Basque, Catalan, Italian and Spanish WordNets
  - Semantic preferences
    - Acquired automatically from Semcor and BNC
  - eXtended WordNet
  - Instances, including named entities

Multilingual Central Repository (MCR)

- In this work, we have used:
  - WN1.6: English WordNet 1.6 synsets and relations
  - WN2.0: English WordNet 2.0 relations (mapped to WN1.6 synsets)
  - XNET: eXtended WordNet (gold, silver and normal)
  - sPref: Selectional preferences
  - sCooc: Coocurrence
  - WN1.7: English WordNet 1.7 synsets and relations
- sPref and sCooc extracted from Semcor
  - system benefits from supervised information when using these

Multilingual Central Repository (MCR)

- We have tried different set of relations

Name	Relations	#synsets	#relations
M16	WN1.6, REL2.0, XNET, sPref, sCooc	99,634	1,651,445
M16_wout_sPref	WN1.6, REL2.0, XNET, sCooc	99,634	1,519,833
M16_wout_sCooc	WN1.6, REL2.0, XNET, sPref	99,632	798,453
M16_wout_Xnet	WN1.6, REL2.0, sPref, sCooc	99,238	1,169,300
M16_wout_Semcor	WN1.6, REL2.0, XNET	99,632	637,290
M17	WN1.7, XNET	109,359	620,396
M16_wout_WXnet	sPref, sCooc	27,336	1,024,698

- Two main groups
  - M16: Based on WordNet 1.6
  - M17: Based on WordNet 1.7

Outline

Experiment setting

- Applied to Senseval 3 All Words dataset
  - Based on WordNet 1.7
- Contexts of at least 20 words
  - Adding sentences immediately before and after

Experiment results

Relations	All	Noun	Verb	Adj.	Adv.
Semi supervised					
M16	57.30	62.30	49.00	**62.40**	92.90
M16_wout_sPref	**57.90**	**63.10**	**49.80**	61.80	92.90
M16_wout_sCooc	53.00	58.10	44.20	58.30	92.90
M16_wout_Xnet	57.60	**63.10**	49.60	61.00	92.90
M16_wout_WXnet	55.30	58.70	48.70	60.80	85.70
Unsupervised					
M16_wout_semcor	53.70	59.50	45.00	57.80	92.90
M17	**56.20**	**61.60**	**47.30**	**61.80**	92.90

- Supervised relations achieve best overall results
  - Specially sCooc, not so with sPref
  - Using only supervised also yields good results
- Unsupervised results: M17 performs best
  - probably due to mapping noise

Comparison to related work

System	All	Noun	Verb	Adj.	Adv.
Mih05	52.2	-	-	-	-
Sin07	52.4	60.45	40.57	54.14	100
Nav07	-	61.9	36.1	62.8	-
M17	56.20	61.60	47.30	61.80	92.90
MFS	60.9-62.4	-	-	-	-
GAMBL	65.1	-	-	-	-

- *Mih05, Sin07*: create a complete weighted graph with synsets of the words in the input context. Weights calculated with similarity measures. Apply PageRank for disambiguating.
- *Nav07*: create subgraph of LKB using DFS search. LKB: Manually enriched WordNet.

Outline

Conclusions

- Graph-based method for performing knowledge-based WSD
- Exploits the structural properties of the graph underlying the chosen knowledge base
- The method is not tied to any particular knowledge base
- Evaluation performed on Senseval-3 All Words
- Evaluation of separate and combined performance of each type of relation in the MCR
  - Validate the contents of the MCR and their potential for WSD
- MCR valuable for performing WSD
  - Relations coming from hand-tagged corpora are the most valuable
- Version of WordNet is highly relevant
- Our graph-based WSD system is competitive with the current state-of-the-art
  - Yields best results that can be obtained using publicly available data