(Un)supervised Approaches to Metonymy Recognition

Yves Peirsman

University of Leuven
RU Quantitative Lexicology and Variational Linguistics

Overview

1. Introduction
2. Unsupervised Metonymy Recognition
3. Supervised Metonymy Recognition
4. Conclusions and outlook

Overview

1. Introduction
2. Unsupervised Metonymy Recognition
3. Supervised Metonymy Recognition
4. Conclusions and outlook

1. Introduction: Metonymy

- A word which does not refer to its original referent A, but to a referent B that is contiguously related to A.
- Metonymical patterns:
  - place for people: *Brussels* opposed to the decision.
  - author for work: He likes to read *Hemingway*.
  - organization for product: He drives a *Mercedes*.
- Ubiquity in everyday language
- Metonymy resolution = recognition + interpretation

1. Introduction: Metonymy recognition

- Metonymy recognition as a subproblem of WSD:
  - possibly metonymical words are polysemous words
  - the recognition of a metonymy = automatic assignment of a sense label to a polysemous word
  - Major advantages:
    - corpus-based machine learning algorithms from WSD
    - model for an entire semantic class of words (e.g. all country names)

1. Introduction: Metonymy recognition

- Metonymy recognition as WSD: Markert and Nissim
  - 1,000 country names, organization names from the BNC
  - WSD algorithms (decision tree, Naïve Bayes)
  - combination of
    - grammatical information: role and head
    - semantic information: Dekang Lin’s (1998) thesaurus
  - promising results
    - Acc = 87% for the country names
    - Acc = 76% for the organization names

1. Introduction: Metonymy recognition

- Main disadvantage:
  - Need for manual annotation:
    hinders development of wide-scale metonymy recognition system

- 2 possible solutions:
  - Unsupervised Learning: no labelled training examples
  - Active Learning: only informative training examples are labelled

Overview

1. Introduction
2. Unsupervised Metonymy Recognition
    2.1 Word Sense Discrimination
    2.2 Experiments
    2.3 Discussion
3. Supervised Metonymy Recognition
4. Conclusions and outlook

2.1 Word Sense Discrimination

- Schütze (1998) clusters similar contexts of words
- Algorithm:
  - word vectors for all words in corpus with co-occurrence info
  - context vectors: sum of word vectors of target co-occurrences
  - optional dimensionality reduction
  - clustering
  - clusters are assumed to correspond to target senses
- Classification:
  - a test instance is assigned to the cluster with the closest centroid

2.1 Word Sense Discrimination

- Metonymy is different from the ambiguous targets that WSD usually addresses:
  - WSD is very good at distinguishing between text topics.
    e.g. *plant* ~ biology vs economy

    => relevant for metonymy?
    e.g. *Brussels* ~? politics vs tourism
  - Instead, for metonymy, grammatical relations seem to play a bigger role.

2.1 Word Sense Discrimination

- Four research questions:
  - Can the algorithm be applied to a set of mixed target words?
    - This will result in more co-occurrences.
  - What context size should we select?
    - Big for text topics, small for grammatical relations?
  - Is SVD a good idea?
    - Word dimensions or topical dimensions?
  - Should features be selected on the basis of a statistical test?

2.2 Experiments

- First round: Hungary data

	+LL, +SVD		+LL, -SVD		-LL, +SVD		-LL, -SVD	
	Acc	F	Acc	F	Acc	F	Acc	F
15	55.73	37.74 ***	63.16	30.14	58.93	33.22*	59.86	31.63
12	56.45	37.39 ***	57.89	32.23	58.41	36.13**	57.69	31.67
10	58.31	37.07 ***	58.72	35.28**	56.45	34.06*	58.20	34.57*
7	55.01	37.89 ***	55.01	36.44**	56.35	37.70***	64.50	34.85***
5	55.01	26.85	55.62	23.49	63.78	36.30***	65.12	33.98**

  – Algorithm never beats the majority baseline.
  – However, many times the proportion of metonymies in one output cluster is significantly larger than that in the other. The system thus regularly beats this random baseline.
  * : $p < 0.05$ ** : $p < 0.01$ *** : $p < 0.001$

2.2 Experiments

- First round: Hungary data

	+LL, +SVD		+LL, -SVD		-LL, +SVD		-LL, -SVD	
	Acc	F	Acc	F	Acc	F	Acc	F
15	55.73	37.74 ***	63.16	30.14	58.93	33.22*	59.86	31.63
12	56.45	37.39 ***	57.89	32.23	58.41	36.13**	57.69	31.67
10	58.31	37.07 ***	58.72	35.28**	56.45	34.06*	58.20	34.57*
7	55.01	37.89 ***	55.01	36.44**	56.35	37.70***	64.50	34.85***
5	55.01	26.85	55.62	23.49	63.78	36.30***	65.12	33.98**

  - performance is better with SVD
  - smaller context windows lead to better results
  - with larger contexts, statistical feature selection is helpful

2.2 Experiments

- Countries

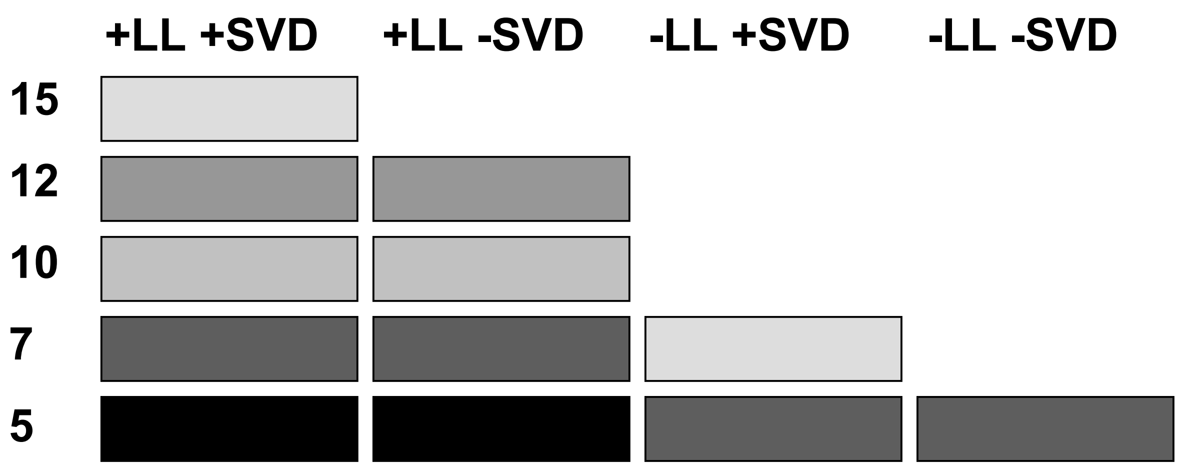
	+LL, +SVD		+LL, -SVD		-LL, +SVD		-LL, -SVD	
	Acc	F	Acc	F	Acc	F	Acc	F
15	57.47	29.76	58.35	24.35	58.79	24.55	60.00	23.21
12	59.23	33.63***	58.68	33.57***	64.95	24.23	63.74	21.43
10	55.71	32.04*	59.34	38.33***	62.31	20.79	66.70	18.77
7	58.90	35.74***	62.75	35.43***	61.32	25.42	67.69	26.13
5	60.55	32.90***	67.14	36.25***	62.31	31.26**	68.46	30.51**

2.2 Experiments

- Organizations

	+LL, +SVD		+LL, -SVD		-LL, +SVD		-LL, -SVD	
	Acc	F	Acc	F	Acc	F	Acc	F
15	54.37	45.35**	53.79	40.80	52.23	34.75	55.92	36.59
12	53.98	43.84*	55.34	42.21*	53.79	32.39	56.21	32.59
10	54.17	44.34**	53.30	41.98	52.33	35.82	54.85	35.15
7	55.44	44.63**	55.24	43.99**	54.66	44.07**	54.76	41.01
5	59.61	42.54***	58.93	44.27***	56.80	46.19***	55.34	46.01***

2.2 Experiments



- smaller contexts are preferable
- statistical feature selection absolutely necessary with mixed sets
- influence of SVD unclear

2.3 Discussion

- Problems:
  - limited amount of data
  - ignorance of syntax: ‘bag-of-word’ approach
- Bag-of-word approaches
  - Original proponents of vector-based techniques (Landauer & co) believe ignorance of syntax is a good thing.
  - Others believe this is a bad thing:
    - Wiemer-Hastings & Zipitria: structured representation of sentences leads to better results
    - Sahlgren: “plea for linguistics”

2.3 Discussion

- Disappointing results for metonymy recognition suggest that some kinds of polysemy may indeed require syntactic information.
- Unsupervised metonymy recognition has so far been relatively unsuccessful.
- Possible solution: combination of
  - Memory-Based Learning
  - Active Learning

Overview

1. Introduction
2. Unsupervised Metonymy Recognition
3. Supervised Metonymy Recognition
    3.1 Memory-Based Learning
    3.2 Active Learning
    3.3 Discussion
4. Conclusions and outlook

3.1. Memory-Based Learning

- MBL = simple “lazy” learning algorithm:
  - in the training stage, all training examples are stored
  - in the test stage, the classifier
    - compares the test instance to all training examples
    - finds the most similar training instances
    - and assigns their most frequent label
- Even without semantic information, this achieves results similar to Nissim and Markert’s
  - countries: role, head, pres of 2nd head, 2nd head
    Acc = 87%
  - organizations: role, head, number, determiner, # roles
    Acc = 75%

3.2 Active Learning

- Active Learning algorithms
  - select those examples that are most interesting for manual annotation
  - uninteresting examples are left unlabelled

    $\Rightarrow$ reduction of effort
  - start with a small number of annotated seed instances and a large pool of unlabelled instances

3.2 Active Learning

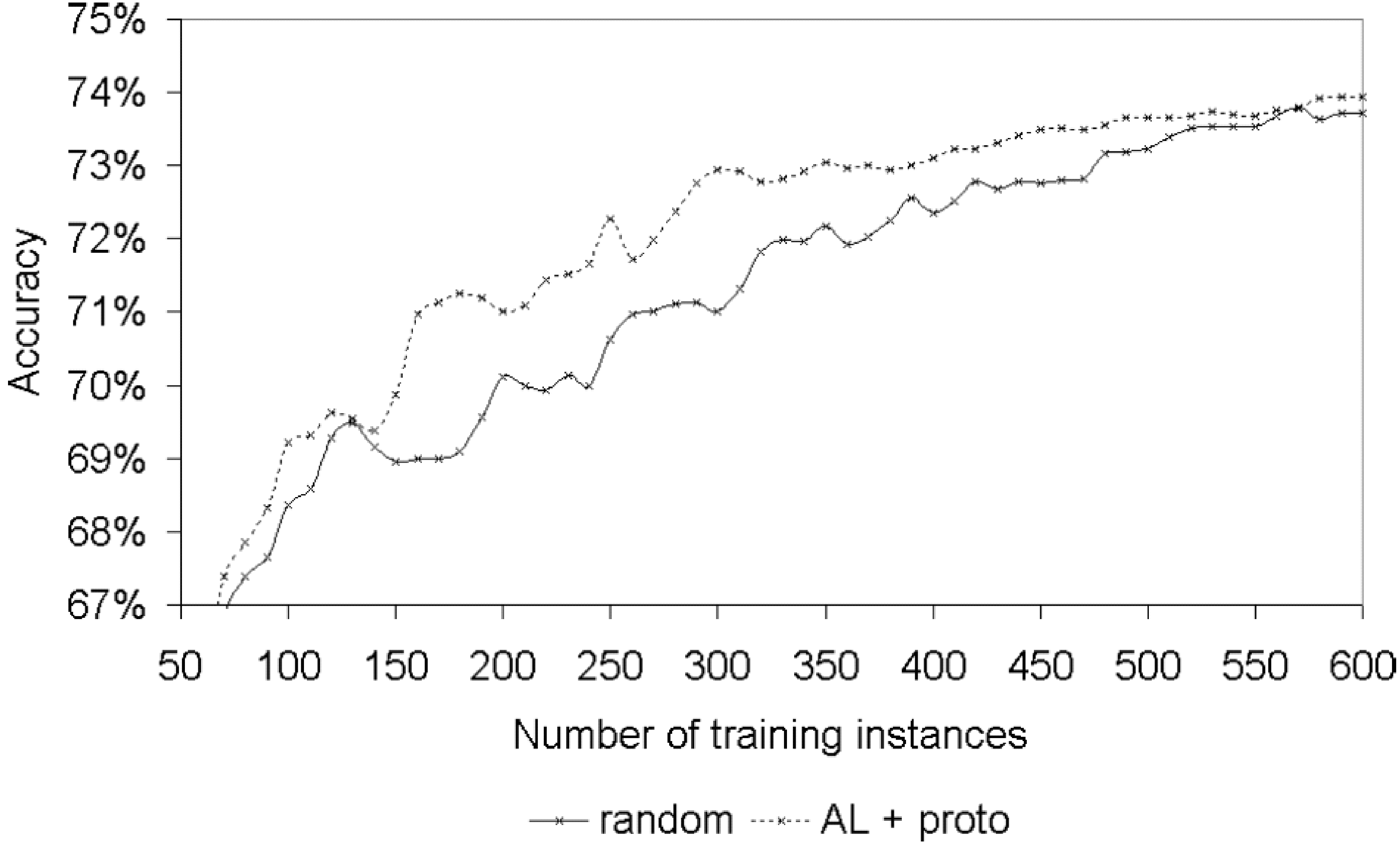
- Algorithm:
  - Classifier trains on the seeds,
  - and labels the unlabelled instances.
  - Algorithm selects instances whose classification is most uncertain,
  - and adds those to the labelled data.
  - Repeat

- “Uncertainty?”
  - e.g., most likely label has low probability (“lowest best P”)

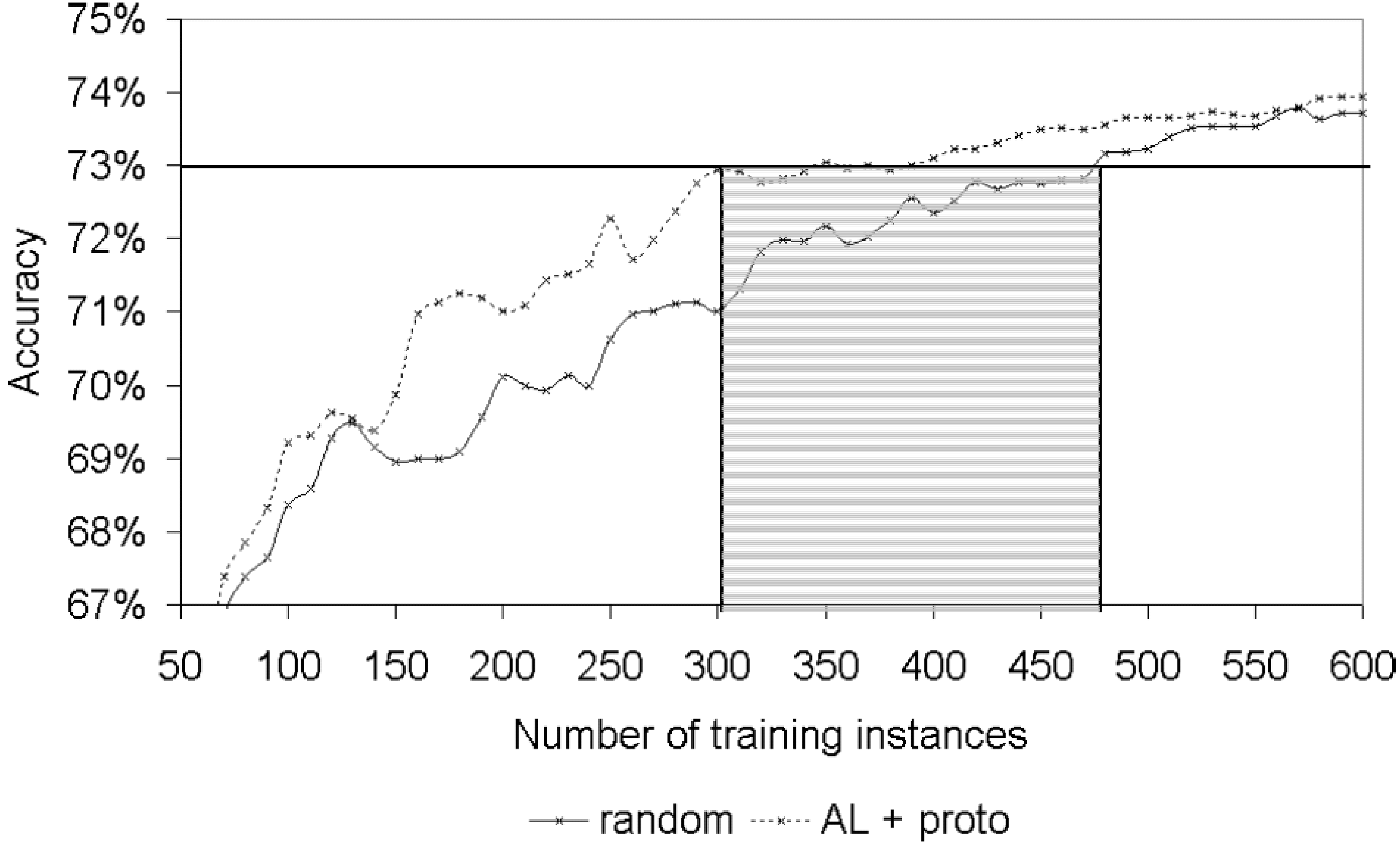
3.2 Active Learning

- Problem: MBL gives distances, not probabilities.
- Hypothesis: distances can be interpreted as a measure of certainty, too.
- Distance-based algorithm:
  - selects “prototypical seeds”
  - on each round, it picks a mix of “distant” and “close” training instances:
    - large distance guarantees new information
    - closer instances are added in order to make labelled set more representative of entire data set
- Experiments on country and organization data

3.2 Active Learning



3.2 Active Learning



3.3 Discussion

- Algorithms should be tested on other data sets.
- Promising results:
  - reductions of 30%
  - limited number of features affects
    - calculation of prototypicality
    - estimation of certainty
  - restricted to Markert and Nissim's data
- Construction of a wide-scale metonymy recognition algorithm need not require an unrealistic amount of annotated training data.

Overview

1. Introduction
2. Unsupervised Metonymy Recognition
3. Supervised Metonymy Recognition
4. Conclusions and outlook

4. Conclusions & Outlook

- Knowledge acquisition bottleneck in metonymy recognition
- Unsupervised metonymy recognition: not yet successful
  - lack of syntax?
- Good results by a combination of
  - Memory-Based Learning: “lazy” learning algorithm
  - Active Learning: selection of informative data
- Future work:
  - new classes of data
  - more informative features
  - bigger unlabelled training pool

for further information:

http://wwwling.arts.kuleuven.be/qlvl
http://perswww.kuleuven.be/yves_peirsman
yves.peirsman@arts.kuleuven.be