Domain Adapted Word Sense Disambiguation

Stefan Gorzitze, Jonas Placzek, Tobias Kostyra
Department of Computational Linguistics
Heidelberg University, Germany

I. Introduction/ Motivation

- **Word Sense Disambiguation:**
  - Distinguishing between the meanings of words in context.
- Used in many NLP applications (i.e. Machine translation, Q-A systems etc.).
- **The problem:** common WSD algorithms only produce satisfying results when used on the same domain they are trained on.

Goal: develop a WSD system with included domain adaptation.

- **Two approaches (Figure 1):**
  1. *Supervised,* via machine learning, the main approach.
  2. *Unsupervised,* via a graph structure from the UKB tool, to evaluate our results against.
- The disambiguation task concentrated on nouns, verbs and adjectives.
- We used the coarse grained WordNet SuperSense classes (Figure 2).

Testing was done on three domains: the SemCor Corpus as base corpus, a collection of ritual texts and recipes.

- The disambiguation process is preceded by a preprocessing step to prepare the input texts for the algorithm.

II. The supervised approach

- **General info:**
  - System „learns“ to correctly label of senses by manually annotated data (from both source and target domain).
  - Extracting of relevant features.
  - Machine learning algorithm: Naive Bayes (Weka).
- **Used training data:**
  - Source domain (huge data amount): SemCor corpus (~200.000 annotated words).
  - Target domains (little data amount): self-annotated data for both domains (~130 sentences each).
- **Features (Figure 3):**
  - No syntantic features due to sentence structur of target domains (no parser applicable).
  - Avoidance of too many features (feature overfitting).

	Used words	Window size
lemmata	nouns, verbs	w-2 | w+2
word types	nouns, verbs	w-2 | w+2
word senses	nouns, verbs	w-2
POS tags	all words	w-2 | w+2

Figure 3: Overview of used features.

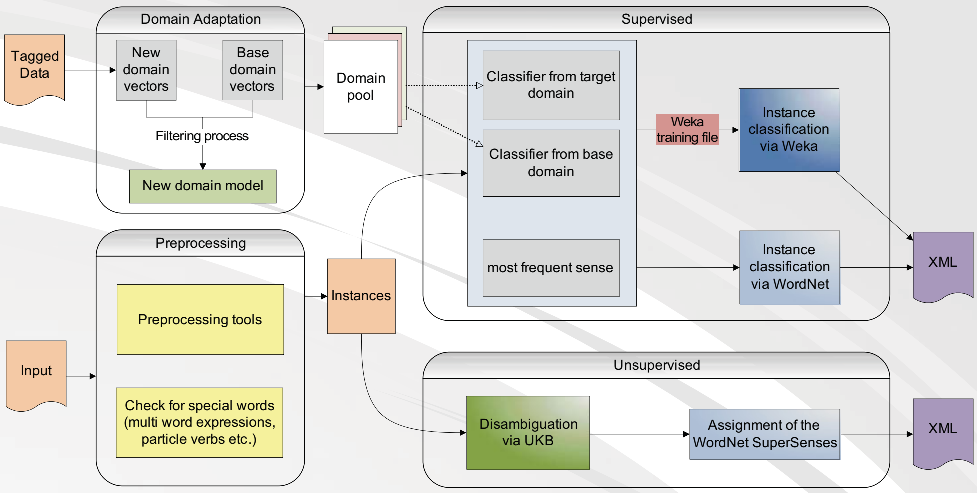

Figure 1: The algorithm.

adj.all	noun.possession
adj.pert	noun.process
adj.ppl	noun.quantity
adv.all	noun.relation
noun.Tops	noun.shape
noun.act	noun.state
noun.animal	noun.substance
noun.artifact	noun.time
noun.attribute	verb.body
noun.body	verb.change
noun.cognition	verb.cognition
noun.communication	verb.communication
noun.event	verb.competition
noun.feeling	verb.consumption
noun.food	verb.contact
noun.group	verb.creation
noun.location	verb.emotion
noun.motive	verb.motion
noun.object	verb.perception
noun.other	verb.possession
noun.person	verb.social
noun.phenomenon	verb.stative
noun.plant	verb.weather

Figure 2: All 46 WordNet SuperSenses.

	domain	supervised			unsupervised		
		noun	verb	adj	noun	verb	adj
all words	base	81	77	99	78	75	99
	ritual	70	65	98	68	65	97
	recipe	92	66	100	89	68	100
polysemous words only	base	65	59	99	69	57	99
	ritual	58	58	97	47	35	98
	recipe	81	56	100	84	66	100

Figure 4: Evaluation results (F-measure).

- **Domain Adaptation:**
  - Adaptation of trained instances of source domain for target domains.
  - Exclusion of instances with similar feature vectors but different senses
  - Jaccard coefficient for calculation of vector similarity.

III. The unsupervised approach

- Uses the UKB tool to disambiguate building a graph around the data of the contexts.
- Implemented to evaluate against the supervised approach.
- The algorithm wraps the preprocessed data and feeds it to the UKB tool for disambiguation.
- The processed data use SenseIDs to get their SuperSenses directly from WordNet.
- The SuperSenses get mapped on the input data and an xml file is created, containing all the disambiguated senses and their char positions in the input text.

IV. Evaluation

- Evaluation was done using manually annotated data from the three test domains (Figure 4).
- Adjectives got best results, due to the fact that WordNet provides only three SuperSenses for Adjectives.
- The results with monosemous words are remarkably better.
- An experiment to broaden the number of senses by using finer grained senses produced notably worse results.

V. References

- Ixa Group: *UKB: Graph Based Word Sense Disambiguation and Similarity.* Basque Country, http://ixa2.si.ehu.es/ukb.
- Daumé, H. III: *Frustratingly Easy Domain Adaption.* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, June 2007: 256–263.
- Reiter, N. et al.: *Adapting Standard NLP Tools and Resources to the Processing of Ritual Descriptions.* Proceedings of ECAI 2010 workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Lisbon, August 2010: 39-46.
- Haas, M., Schamoni, H., Wittl, T., Zeller, B.: *RECIPE.* Software project, Heidelberg, winter term 2009/ 2010.