Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution

Detlef Prescher, Stefan Riezler, Mats Rooth

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Germany

August 4, 2000

Overview

- Motivation
- Disambiguation with probabilistic class-based lexicons
  - Induction of probabilistic class-based lexicons via clustering
  - Use of class-based frequencies for disambiguation via lexicon look-up
- Evaluation
  - Evaluation on a pseudo-disambiguation task
  - Evaluation on a real-world bilingual corpus
- Conclusion

Motivation

Aim: Resolution of lexical ambiguities in applications such as machine translation or lexicalized parsing, where:

- some parsing is likely to be done,
- but no large amount of contextual information is available.

Method:

- Induce a probabilistic class-based lexicon providing **frequency information** and **class-based probabilities** about verbs and their arguments.
- Resolve lexical ambiguities by a simple look-up in this probabilistic class-based lexicon.

Clustering

Given: Data tuples (v, n) of verbs v and their noun arguments n.

Aim: Smooth class-based probability model on the data.

Solution: **Clustering model**

$$p(v, n) = \sum_{c \in C} p(c, v, n) = \sum_{c \in C} p(c) \cdot p(v|c) \cdot p(n|c)$$

parameterized by $p(c), p(v|c), p(n|c)$ and re-estimated via the EM-algorithm.

Reference: Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of ACL'99*, College Park, MD.

Example: Class 19 - “locative action”

Class 19 PROB 0.0235		0.0250	0.0211	0.0125	0.0111	0.0106	0.0096	0.0085	0.0084	0.0081	0.0079	0.0076	0.0068	0.0067	0.0061	0.0059	0.0056	0.0053	0.0050	0.0049	0.0049	0.0048	0.0048	0.0047	0.0047	0.0046	0.0046	0.0045	0.0044	0.0043	0.0043
		room	area	world	meeting	range	school	service	building	road	street	market	course	doctor	place	part	mind	class	scene	path	group	work	gray	programme	line	life	garden	body	miles	system	period
0.0629	enter	●	●	●		●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●		●	●	●		●	●
0.0386	cover		●	●	●	●	●	●		●	●	●	●	●	●	●			●		●	●	●	●	●			●	●		●
0.0321	call	●	●	●	●		●	●		●	●				●	●	●	●	●	●	●	●	●	●	●	●				●	●
0.0236	include	●	●	●	●	●	●	●	●		●		●	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●
0.0226	run	●		●			●	●					●	●	●	●		●		●	●	●		●	●	●			●	●	
0.0214	attend				●		●	●		●		●		●	●	●		●	●		●	●		●							
0.0173	cross	●	●			●			●	●	●		●			●	●			●			●		●		●	●			
0.0136	dominate	●	●	●	●			●	●	●	●	●			●			●	●			●	●			●		●		●	●
0.0132	have	●		●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●
0.0126	attract		●		●		●	●		●			●		●	●			●		●	●		●					●	●	●
0.0124	occupy	●	●	●		●					●			●	●	●	●						●					●	●		

Induction of a Probabilistic Class-Based Lexicon (ProbLex)

Given: Clustering model $p_{\text{cluster}}(.)$; samples of nominal fillers of the object slots of fixed transitive verbs.

Aim: Probabilistic class-based models on nominal fillers of fixed verbs, using the classes of the given clustering model $p_{\text{cluster}}(.)$.

Solution: **ProbLex model:**

$$p(n) = \sum_{c \in C} p(c, n) = \sum_{c \in C} p(c) \cdot p_{\text{cluster}}(n|c)$$

parametrized by $p(c)$ and re-estimated by EM (along the lines of Rooth et al., 1999)

Example: Two Entries of ProbLex - $cross$ and $mobilize$

cross 19	0.692	mobilize 6	0.386
mind	74.2	force	2.00
road	30.3	people	1.95
line	28.1	army	1.46
bridge	27.5	sector	0.90
room	20.5	society	0.90
border	17.8	worker	0.90
boundary	16.2	member	0.88
river	14.6	company	0.86
street	11.5	majority	0.85
atlantic	9.9	party	0.80

Estimated class-based frequencies of noun n in the sample $f(.)$ of objects of a fixed verb:

$$f_c(n) = f(n) \cdot p(c|n), \quad c \in C$$

Disambiguation with ProbLex

Given: Applications such as machine translation or lexicalized parsing, where some parsing is likely to be done, but no large context-windows are available.

Aim: Find a correct target-noun and use only minimal context information.

Solution: Resolve noun ambiguities $\mathcal{A}$ by using **frequency information** and **class-based probabilities** provided by the governing target-verb.

Simple look-up in ProbLex: choose target noun $\hat{n}$ (and a class $\hat{c}$) such that

$$f_{\hat{c}}(\hat{n}) = \mathrm{max}_{n \in \mathcal{A}, c \in C} \; f_c(n)$$

where $f_c(n)$ is calculated by the frequency $f(n)$ of n in the **combined sample of objects and translation alternatives**.

Examples for Target-Word Ambiguities

(ID 160867)

Es gibt einige alte Passvorschriften, die besagen, dass man einen Pass haben muss, wenn man die **Grenze** *überschreitet.*

There are some old provisions regarding passports which state that people crossing the **{border/ frontier/ boundary/ limit/ periphery/ edge}** should have their passport on them.

(ID 201946)

Es gibt schliesslich keine Lösung ohne die Mobilisierung der bürgerlichen **Gesellschaft** *und die Solidarität der Demokraten in der ganzen Welt.*

There can be no solution, finally, unless civilian **{company/ society/ companionship/ party/ associate}** is mobilized and solidarity demonstrated by democrats throughout the world.

Evaluation on Pseudo-Disambiguation Task for Noun-Ambiguity

Pseudo-disambiguation task: An observed verb-noun pair (v, n) should be preferred over a completely random pair (v, n').

Generalization power: For a test set of 298 (v, n, n') types, precision was calculated as the number of times the disambiguation decided for the non-random target-noun (Clustering: $p_{\text{cluster}}(v, n) \geq p_{\text{cluster}}(v, n')$; ProbLex: $\hat{n} = n$).

Results:

ambiguity	human baseline	clustering	ProbLex
2	73.5 %	79.0 %	88.3 %

Evaluation on a Real-World Bilingual Corpus

- Resources:
  - **Sentence-aligned bilingual corpus** of 9 million tokens for German and English (mlcc = multilingual corpus for cooperation)
  - Word-to-word translations from **online-available dictionaries**

- Semi-automatic filtering and manual mark-up resulted in [1]:
  - **3 dictionaries**, differing in the average number of translations (8.63, 2.83, 2), and the semantic distance of translations ("all translations", "semantically distinct", "polysemous")
  - **Associated evaluation corpora**, differing in size (1340, 814, 219 sent.)
  - For each bilingual sentence, an **annotation** (n_G, n_E, v_E) consisting of: n_G the German source-noun, n_E the correct English target-noun, and v_E the English governing verb

[1] Available on http://www.ims.uni-stuttgart.de/projekte/gramotron

Dictionaries Extracted from Online Resources

Art	**form, type, way**, fashion, fit, kind, wise, manner, species, mode, sort, variety
Aufgabe	**abandonment, office, task**, exercise, lesson, giveup, job , problem, tax
Begriff	concept, item, notion, idea
Boden	**ground, land, soil**, floor, bottom
Fehler	**error, shortcoming**, blemish, blunder, bug, defect, demerit, failure, fault, flaw, mistake, trouble, slip, blooper, lapsus
Geschichte	**history, story**, tale, saga, strip
Gesellschaft	**company, society**, companionship, party, associate
Grenze	**border, frontier**, boundary, **limit**, periphery, edge
Karte	**card, map**, ticket, chart
Lage	site, situation, position, bearing, layer, tier
Mangel	deficiency, lack, privation, want, shortage, shortcoming, absence, dearth, demerit, desideratum, insufficiency, paucity, scarceness
Menge	**amount, deal, lot**, mass, multitude, plenty, quantity, quiverful, volume, abundance, aplenty, assemblage , crowd, batch, crop, heap, lashings, scores, set, loads, bulk
Seite	**page, party, side**, point, aspect
Stimme	**voice, vote**, tones
Termin	**date, deadline, meeting**, appointment, time, term
Verbindung	**association, contact, link**, chain, conjunction, connection, fusion, joint , compound, alliance, catenation, tie, union, bond, interface, liaison, touch, relation, incorporation
Verbot	ban, interdiction, prohibition, forbiddance
Wahl	**election, option**, choice , ballot, alternative, poll , list
Weg	**path, road, way**, alley, route, lane
Widerstand	resistance, opposition, drag
Zeichen	character, icon, sign, signal, symbol, mark, token, figure, omen
Ziel	**aim, destination, end**, designation, target, goal, object, objective, sightings, intention, prompt

Disambiguation Results for Clustering versus Probabilistic Lexicon Methods

ambiguity	random	major sense	empirical distrib.	clustering	ProbLex
8.63	14.2 %	31.9 %	P: 46.1 % E: 36.2 %	43.3 %	49.4 %
2.83	35.9 %	45.5 %	P: 60.8 % E: 49.4 %	61.5 %	68.2 %

Precision: P = #correct / (#correct + #incorrect)
Effectiveness: E = #correct / (#correct + #incorrect + #don't know)

- Clustering solves "don't know" problems of empirical distribution
- ProbLex outperforms clustering using estimated class-based frequencies

Precision for Finding Correct and Acceptable Translations Using ProbLex

source	target	correct	accept.
Seite	page side	76.9 %	76.9 %
Sicherheit	guarantee safety	93.8 %	93.0 %
Verbindung	connection link	58.8 %	93.8 %
Verpflichtung	commitment obligation	73.2 %	94.1 %
Ziel	objective target	72.5 %	85.5 %
overall precision		78 %	90 %

Comparison of Unsupervised Lexical Disambiguation Methods

disambiguation method	corpus size	ambiguity	random	precision	random (standardized)	precision (standardized)
ProbLex	1 340	8.63	14.2 %	49.4 %	53.4 %	79.7 %
	814	2.83	35.9 %	68.2 %	50.5 %	77.5 %
	219	2	50.0 %	78.0 %	50.0 %	78.0 %
Dagan, Itai 94	103	2.27	44.1 %	P: 91.4 % E: 62.1 %	50.0 %	P: 92.7 % E: 66.8 %
Resnik 97	88	3.51	28.5 %	44.3 %	50.0 %	63.8 %
SENSEVAL 00	2 756	9.17	10.9 %	P: 20-65 % E: 3-54 %	50.0 %	P: 60-87 % E: 33-83 %
Yarowsky 95	$\sim$ 37 000	2	50.0 %	96.5 %	50.0 %	96.5 %
Schütze 92	$\sim$ 3 000	2	50.0 %	92.0 %	50.0 %	92.0 %

Standardization: $p^{1/\log_2 amb}$ for precision p and ambiguity factor amb

Conclusion

- Our disambiguation method is restricted to the information provided by a **simple look-up in a probabilistic class-based lexicon**

- **Disambiguation based on probabilistic lexica performed satisfactory on** randomly selected examples from real-world bilingual corpora which constitutes **a realistic hard task**

- **Hybrid models** combining frequency information and class-based probabilities **outperform both pure frequency-based models and pure clustering models**