Inducing a Semantically Annotated Lexicon via EM-Based Clustering

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, Franz Beil

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Germany

June 23, 1999

Overview

- Outline
- Inducing selection classes via EM-based clustering
  - Theory
  - Clustering examples
  - Evaluation on pseudo-disambiguation task
- Inducing class labels for subcat-slots via EM-estimation
  - Theory
  - Experiments with English and German data
  - Linguistic evaluation
- Conclusion

Outline

Aim:

Induction of *semantic class labels for slots of subcategorization frames* from *unannotated* corpora, i.e., no taxonomy of semantic classes required, rather semantic classes are tailored to unannotated data from arbitrary languages.

Method:

- Collect data tuples consisting of verbs annotated with subcat-frames and their nominal fillers by parsing large corpora,
- Induce semantic selection classes for verb-frames and nouns via EM-based clustering,
- Induce class labels for subcategorization slots of given verb-frames by further application of EM to noun-filler data.

Inducing Selection Classes via EM-Based Clustering

Given: Data tuples $\langle v, n \rangle$ of intransitive and transitive verb-frames v and their subjects and objects n.

Aim: Smooth class-based probability model on data, together with well-defined and simple method for statistical inference from unannotated data.

Class-based selection model:

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

parameterized by $\theta = \langle \theta_c = p(c), \theta_{v,c} = p(v|c), \theta_{n,c} = p(n|c) \rangle$.

EM-Mapping for Class-Based Selection Model

Let $y = \langle v, n \rangle$ and $x = \langle c, y \rangle$ be fixed, and $f(y)$ be the sample-frequency of y. The **EM-estimates** of $\theta = \langle \theta_c, \theta_{v,c}, \theta_{n,c} \rangle$ are given by the mapping M as

$$
\begin{aligned}
M(\theta_{vc}) &= \frac{\sum_{y \in \{v\} \times N} f(y) p_\theta(x|y)}{\sum_y f(y) p_\theta(x|y)}, \\
M(\theta_{nc}) &= \frac{\sum_{y \in V \times \{n\}} f(y) p_\theta(x|y)}{\sum_y f(y) p_\theta(x|y)}, \\
M(\theta_c) &= \frac{\sum_y f(y) p_\theta(x|y)}{\sum_y f(y)}.
\end{aligned}
$$

According to the EM-theory, each estimation step increases the likelihood $L = \prod_y p_\theta(y)^{f(y)}$; a sequence of estimates converges to a (local) maximum of L.

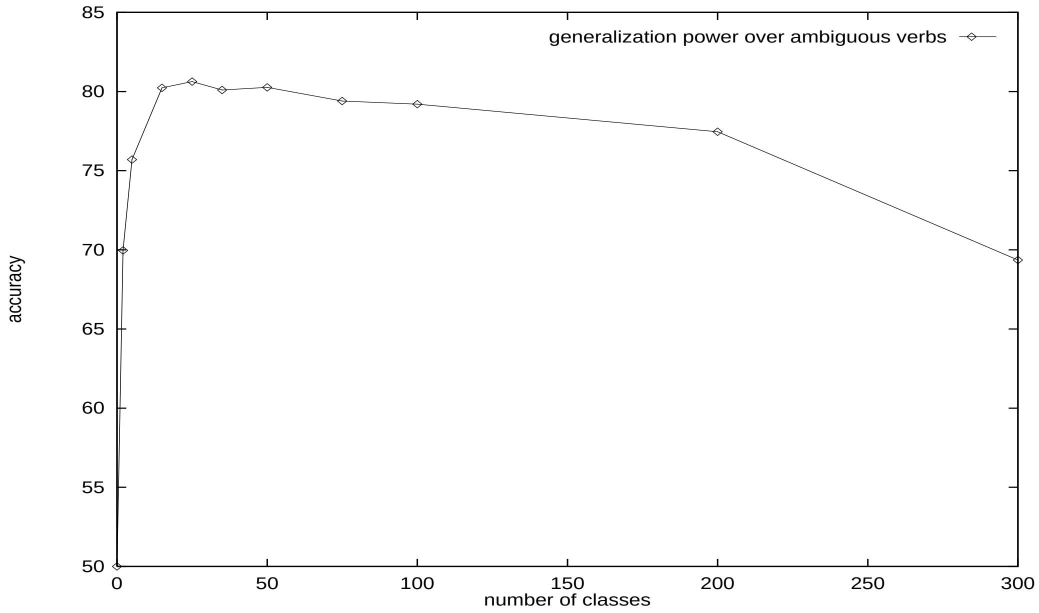
Class 5 PROB 0.0412		0.0148	0.0084	0.0082	0.0078	0.0074	0.0071	0.0054	0.0049	0.0048	0.0047	0.0046	0.0041	0.0040	0.0040	0.0039	0.0039	0.0039	0.0039	0.0038	0.0038	0.0038	0.0037	0.0035	0.0035	0.0035	0.0034	0.0033	0.0033	0.0033	0.0033
		man	ruth	corbett	doctor	woman	athelstan	cranston	benjamin	stephen	adam	girl	laura	maggie	voice	john	harry	emily	one	people	boy	rachel	ashley	jane	caroline	jack	burun	juliet	blanche	helen	edward
0.0542	ask.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●
0.0340	nod.as:s		●	●	●	●	●	●	●	●		●	●	●		●	●	●			●	●	●	●	●	●	●	●	●	●	●
0.0299	think.as:s	●	●	●	●	●	●			●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●		●	●	●	●
0.0287	shake.aso:s	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●		●	●	●	●	●	●	●	●	●		●	●
0.0264	smile.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●		●	●	●	●	●
0.0213	laugh.as:s	●	●		●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
0.0207	reply.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●			●				●	●
0.0167	shrug.as:s	●		●	●	●	●	●	●	●	●	●	●	●			●	●			●	●		●	●	●	●	●	●	●	●
0.0148	wonder.as:s	●	●	●	●	●	●			●	●	●				●	●	●		●	●	●	●	●	●	●	●	●	●	●	●
0.0141	feel.aso:s	●		●	●	●				●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●		●	●	●	●
0.0133	take.aso:s	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●
0.0121	sigh.as:s	●	●	●	●	●	●		●	●	●		●			●	●	●		●		●	●	●	●	●	●		●	●	●
0.0110	watch.aso:s	●	●	●		●	●	●	●	●	●	●		●		●	●	●	●	●	●	●		●		●	●	●		●	
0.0106	ask.aso:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●			●		●			●	●	●
0.0104	tell.aso:s	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●
0.0094	look.as:s	●	●	●	●	●	●	●	●	●	●	●				●		●	●	●	●	●				●	●	●			●
0.0092	give.aso:s	●	●		●	●			●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

Class 17 PROB 0.0265		0.0379	0.0315	0.0313	0.0249	0.0164	0.0143	0.0110	0.0109	0.0105	0.0103	0.0099	0.0091	0.0089	0.0088	0.0082	0.0077	0.0073	0.0071	0.0071	0.0070	0.0068	0.0067	0.0065	0.0065	0.0058	0.0057	0.0057	0.0054	0.0051	0.0050
		number	rate	price	cost	level	amount	sale	value	interest	demand	chance	standard	share	risk	profit	pressure	income	performance	benefit	size	population	proportion	temperature	tax	fee	time	power	quality	supplely	money
0.0437	increase.as:s	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
0.0392	increase.aso:o		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●		●	●	●	●	●
0.0344	fall.as:s	●	●	●	●	●	●	●	●		●	●	●	●		●	●	●	●			●	●	●	●		●		●	●	
0.0337	pay.aso:o	●	●	●		●	●		●	●	●			●				●	●	●			●			●	●				●
0.0329	reduce.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
0.0257	rise.as:s	●	●	●	●	●	●	●	●		●		●	●	●	●	●	●		●	●	●	●	●	●		●	●	●	●	
0.0196	exceed.aso:o	●	●	●	●	●	●		●	●	●	●	●			●		●		●	●	●	●		●	●	●	●	●	●	
0.0177	exceed.aso:s	●	●	●	●	●	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●		●			●	●
0.0169	affect.aso:o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●		●	●	●	●	●	
0.0156	grow.as:s	●						●	●	●	●			●	●	●	●	●		●	●	●	●					●			●
0.0134	include.aso:s	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●		●	●			●	●	●	●	●	●
0.0129	reach.aso:s	●	●	●	●	●	●	●		●	●			●		●	●		●	●		●	●	●			●	●	●	●	●
0.0120	decline.as:s	●	●	●	●	●		●	●	●	●		●	●	●	●					●	●	●	●					●	●	
0.0102	lose.aso:o	●	●		●		●	●	●	●		●	●	●	●	●	●	●		●		●	●	●		●	●	●	●		●
0.0099	act.aso:s		●	●				●					●							●					●			●		●	
0.0099	improve.aso:o	●	●			●			●			●	●	●		●			●								●	●	●		
0.0088	include.aso:o	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

Evaluation of Clustering Models

Pseudo-disambiguation task: In generalization over unseen pairs, a "natural" pair $\langle v, n \rangle$ should be preferred over a completely random pair $\langle v', n \rangle$.

Generalization power: $\text{accuracy} = \dfrac{|\{\langle v, n, v' \rangle \in \mathit{Eval} : p(n|v) \geq p(n|v')\}|}{|\mathit{Eval}|}$



Inducing Class Labels for Subcat-Slots using EM-Estimation

Given: Selection Class Model $p_{\mathrm{SC}}(\cdot)$ on $\langle v, n \rangle$ data as before; sample $n_1, n_2, \ldots, n_M$ of nominal fillers of subject slot of fixed intransitive verb.

Aim: Tune class-parameters of probability model on noun data to nominal fillers of fixed verbs.

Class-based slot-labeling model:

$$p(n) = \sum_{c \in C} p(c, n) = \sum_{c \in C} p(c) p_{\mathrm{SC}}(n|c)$$

parameterized by $\theta = \langle \theta_c = p(c) \rangle$.

EM-Mapping for Class-Based Slot Labeling Model

EM-Mapping:

$$M(\theta_c) = \frac{\sum_{n \in N} f(n) p_\theta(c|n)}{\sum_{n \in N} f(n)}$$

According to EM-theory, a sequence of such estimates converges to (local) maximum of likelihood $L = \prod_n p_\theta(n)^{f(n)}$ of sample of subject-fillers.

Output: List of slot labels for fixed verb-frame sorted according to θ_c; List of nominal fillers for slots of fixed verb-frame sorted according to $f(n)p_\theta(c|n)$.

Experiments with English data

Class 5 ("communicative action") labeling subject-slot of intransitive *blush* and *snarl* together with expected noun-fillers:

blush 5	0.982975	*snarl* 5	0.962094
constance	3	mandeville	2
christina	3	jinkwa	2
willie	2.99737	man	1.99859
ronni	2	scott	1.99761
claudia	2	omalley	1.99755
gabriel	2	shamlou	1
maggie	2	angalo	1
bathsheba	2	corbett	1
sarah	2	southgate	1
girl	1.9977	ace	1

Experiments with English data, contd.

Class 17 ("scalar change") labeling subject-slot of intransitive and object-slot of transitive *increase*:

increase 17	0.923698		
number	134.147	proportion	23.8699
demand	30.7322	size	22.8108
pressure	30.5844	rate	20.9593
temperature	25.9691	level	20.7651
cost	23.9431	price	17.9996

increase $(8, 17)$	0.3097650
development - pressure	2.3055
fat - risk	2.11807
communication - awareness	2.04227
supplementation - concentration	1.98918
increase - number	1.80559

Experiments with English data, contd.

Grouping of intransitives taking class 17 as best subject-label:

0.977992 decrease	0.560727 drop
0.948099 double	0.476524 grow
0.923698 increase	0.42842 vary
0.908378 decline	0.365586 improve
0.877338 rise	0.365374 climb
0.876083 soar	0.292716 flow
0.803479 fall	0.280183 cut
0.672409 slow	0.238182 mount
0.583314 diminish	

Comparison to Levin's 48 top-level verb classes shows an agreement with her class of "verbs of change of state", except for last three verbs.
Furthermore, most verbs take class 17 as best intransitive subject-label and as best transitive object-label, pointing at causative alternation.

Conclusion

Proposed algorithms for clustering and slot labeling from unannotated data are

- well-founded in the theory of parametrized statistical inference,
- efficient enough to map very large corpora of observations of frames and their complements to structured lexical entries,
- applicable to any natural language where text samples of sufficient size together with a robust parser are available.

Lexicon Induction for Transitive Verbs

Given: $p_{\mathrm{SC}}(\cdot)$ on $\langle v, n \rangle$ data as before; sample $(n_1, n_2)_1, (n_1, n_2)_2, \ldots, (n_1, n_2)_N$ of nominal fillers of subject and object slot of fixed transitive verb.

Class-based slot labeling model:

$$p(n_1, n_2) = \sum_{c_1, c_2 \in C} p(c_1, c_2, n_1, n_2) = \sum_{c_1, c_2 \in C} p(c_1, c_2) p_{LC}(n_1 | c_1) p_{LC}(n_2 | c_2).$$

EM-mapping for parameters $\theta = \langle \theta_{c_1 c_2} = p(c_1, c_2) \rangle$:

$$M(\theta_{c1c2}) = \frac{\sum_{n_1, n_2 \in N} f(n_1, n_2) p_\theta(c_1, c_2 | n_1, n_2)}{\sum_{n_1, n_2 \in N} f(n_1, n_2)}$$

Experiments with German data

Class 26 labeling subject-slot of intransitive verb "steigen" (rise) and object-slot of transitive verb "senken"(lower):

steigen 26	0.67273	(rise)
Zahl	23.333	(number)
Preis	15.895	(price)
ArbeitsLosigkeit	10.8788	(unemployment)
Lohn	9.72965	(wage)
NachFrage	6.83619	(demand)

senken $(14, 26)$	0.450352	(lower)
BundesBank - LeitZins	5.81457	(Federal bank - base rate)
BundesBank - Zins	2.97838	(Federal bank - interest)
superMarkt - Preis	1	(super market - price)
SommerGeschäft - Verlust	1	(summer business - loss)
BundesBank - DiskontSatz	0.99999	(Federal bank - minimum lending rate)

Experiments with German data, contd.

Grouping of intransitives taking class 26 as best subject-label:

0.741467 ansteigen	(go up)
0.720221 steigen	(rise)
0.693922 absinken	(sink)
0.656021 sinken	(go down)
0.438486 schrumpfen	(shrink)
0.375039 zurückgehen	(decrease)
0.316081 anwachsen	(increase)
0.215156 stagnieren	(stagnate)
0.160317 wachsen	(grow)
0.154633 hinzukommen	(be added)

Comparison to Schuhmacher's verb classification shows correspondence with class of "einfache Änderungsverben" (simple verbs of change of state) except for *anwachsen* (increase) and *stagnieren* (stagnate).