Probabilistic Clustering for Identification of Verb-Noun Collocations

Detlef Prescher, Ulrich Heid

Institute for Natural Language Processing
University of Stuttgart
Germany

March 3, 2000

Overview

- Outline
- Inducing selection classes via probabilistic clustering
  - Probabilistic clustering
  - Examples
- Identification of verb-noun collocations via probabilistic clustering
  - General idea
  - Application to collocations
  - Examples
  - Qualitative evaluation
- Discussion
  - Comparison with Dunning's scores
  - Boosting performance
- Conclusion

Outline

Aim:

Identification of verb-noun collocations from unannotated corpora

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 probabilistic clustering
- Identify verb-noun collocations by comparing the clustering model with the empirical distribution

The lexicographers' view: a wide range of phenomena

Types of phenomena

- "Free" combinations (HAUSMANN 89) and (sortal) selection:
  Französisch | *Deutsch* | *Englisch* | *... sprechen*
- Collocation: *Antrag + stellen*, *Rolle + spielen*, *Risiko + eingehen*
- Funktionsverbgefüge (FVG): *in Bewegung sein* | *geraten* | *bringen*, *Ausdruck finden*; *Ergebnis + liegt vor*, *Interesse + besteht*

⇒ No clearcut dividing lines, but a few tests

Syntactic types

- Verb + object: *Beschluß fassen*
- Verb + subject (cf. MEL'ČUK 92 ETC.):
  Ruhe + kehrt ein, *Frage + kommt auf*, *Problem + besteht*
- Verb + adjunct

Inducing Selection Classes via Probabilistic Clustering

Given: Data tuples (v, n) of 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:

$$\begin{aligned}
p(v,n) &= \sum_{c \in C} p(c,v,n) \\
&= \sum_{c \in C} p(c)p(v,n|c) \\
&= \sum_{c \in C} p(c)p(v|c)p(n|c) \quad \text{(Independence Assumption!)}
\end{aligned}$$

Class 5 PROB 0.0412		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
		0.0148	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
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	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

Identification of Collocations using the Clustering Model

General idea:

- the empirical probability $\hat{p}(v, n)$,
  i.e. the relative frequency of a verb-noun pair (v, n),
  describes the number of occurences of (v, n) in a corpus

- the model probability $p(v, n)$ describes the number of occurences of (v, n) under the specific pre-conditions of the model

$\Rightarrow$ if a (frequent) pair is not conformant to the model pre-conditions, we expect its empirical probability to be much higher than its model probability.

Identification of Collocations using the Clustering Model

Application to identification of collocations: We observe "free coocurrence" of verbs and nouns in the classes of the clustering model. Thus we expect genuine verb-noun collocations (v, n) not to fit into the clustering model.

$\Rightarrow \hat{p} - p$ is big/small for collocations/"free combinations"

$\hat{p} - p$	
7.25709e-05	begehen_VPA.na:NP.AKK Straftat
2.38116e-06	spendieren_VPA.nad:NP.Akk Bier
7.41134e-05	treffen_VPP.n:NP.Nom Entscheidung
2.3708e-06	zerstören_VPP.n:NP.Nom Vegetation
2.73341e-05	schweigen_VPA.n:NP.Nom Waffe
4.77741e-06	spenden_VPA.nad:NP.Nom Franz

$\hat{p} - p$	**TOP-20 Aktiv/OBJ**
0.000344138	haben_VPA.nai:NP.Akk nichts
0.000287041	spielen_VPA.nap:NP.Akk Rolle
0.00021156	lösen_VPA.na:NP.Akk Problem
0.000201252	haben_VPA.nd:NP.Dat Erfolg
0.000165406	machen_VPA.nd:NP.Dat Gebrauch
0.000143906	erreichen_VPA.na:NP.Akk Ziel
0.000126935	haben_VPA.nap:NP.Akk Anspruch
0.000124071	geben_VPA.nap:NP.Akk Problem
0.00011495	machen_VPA.nad:NP.Akk Spass
0.000113718	haben_VPA.nap:NP.Akk Einfluss
0.00011103	geben_VPA.nai:NP.Akk nichts
0.000109662	kosten_VPA.na:NP.Akk Mark
0.000107914	wissen_VPA.na:NP.Akk nichts
0.000106176	haben_VPA.nd:NP.Dat Bestand
0.000104177	leisten_VPA.nap:NP.Akk Beitrag
0.000102838	haben_VPA.nd:NP.Dat Vorrang
9.98574e-05	haben_VPA.nap:NP.Akk Interesse
9.32472e-05	treffen_VPA.na:NP.Akk Entscheidung
8.23121e-05	nachkommen_VPA.nd:NP.Dat Verpflichtung
8.17559e-05	nehmen_VPA.nd:NP.Dat Schaden

$\hat{p} - p$	**TOP-20 Passiv/SUBJ**
0.000144875	lösen_VPP.n:NP.Nom Problem
0.00011534	erfüllen_VPP.n:NP.Nom Forderung
7.94366e-05	lösen_VPP.np:NP.Nom Problem
7.68178e-05	erreichen_VPP.n:NP.Nom Ziel
7.41134e-05	treffen_VPP.n:NP.Nom Entscheidung
6.23074e-05	treffen_VPP.np:NP.Nom Entscheidung
5.60728e-05	erzielen_VPP.np:NP.Nom Einigung
5.53494e-05	besetzen_VPP.n:NP.Nom Stelle
5.16534e-05	ändern_VPP.n:NP.Nom Grundgesetz
4.68707e-05	ausüben_VPP.np:NP.Nom Druck
4.67291e-05	erheben_VPP.n:NP.Nom Anklage
4.46287e-05	tragen_VPP.nd:NP.Nom Rechnung
4.37018e-05	einstellen_VPP.n:NP.Nom Verfahren
4.14156e-05	erfüllen_VPP.n:NP.Nom Bedingung
3.94744e-05	schaffen_VPP.np:NP.Nom Arbeitsplatz
3.94025e-05	schaffen_VPP.n:NP.Nom Voraussetzung
3.74812e-05	erfüllen_VPP.np:NP.Nom Forderung
3.56396e-05	bieten_VPP.nd:NP.Nom Einhalt
3.53789e-05	schaffen_VPP.n:NP.Nom Arbeitsplatz
3.48676e-05	machen_VPP.np:NP.Nom Fehler

$\hat{p} - p$	**TOP-20 Aktiv/SUBJ**
0.000214584	mitteilen_VPA.n:NP.Nom Polizei
0.000129361	bestehen_VPA.n:NP.Nom Gefahr
0.000125135	passieren_VPA.n:NP.Nom etwas
8.14615e-05	ändern_VPA.npr:NP.Nom etwas
8.1289e-05	steigen_VPA.n:NP.Nom Zahl
8.05926e-05	fehlen_VPA.n:NP.Nom Geld
7.79403e-05	mitteilen_VPA.n:NP.Nom Sprecher
6.65567e-05	fallen_VPA.n:NP.Nom Entscheidung
6.41839e-05	ändern_VPA.nr:NP.Nom etwas
6.40753e-05	ausgehen_VPA.np:NP.Nom Gefahr
6.10802e-05	eintreffen_VPA.n:NP.Nom Polizei
6.09915e-05	kommen_VPA.np:NP.Nom Rede
5.93555e-05	kann_VPA.np:NP.Nom Rede
5.72543e-05	vergehen_VPA.nd:NP.Nom sehen
5.70229e-05	sagen_VPA.n:NP.Nom Sprecher
5.66463e-05	melden_VPA.n:NP.Nom Nachrichtenagentur
5.62966e-05	bestehen_VPA.n:NP.Nom Interesse
5.57909e-05	zeigen_VPA.n:NP.Nom Beispiel
5.24866e-05	aufgehen_VPA.n:NP.Nom Rechnung
5.24282e-05	wiederholen_VPA.nr:NP.Nom Geschichte

Qualitative Evaluation: Observations (1/3)

Expected Noun/Verb-Collocations (Examples sorted by noun)

Arbeit aufnehmen_VPA.na:NP.Akk
Arbeit finden_VPA.na:NP.Akk
Arbeit leisten_VPA.na:NP.Akk
Arbeit machen_VPA.na:NP.Akk
Arbeit verrichten_VPA.na:NP.Akk
Verantwortung tragen_VPA.na:NP.Akk
Verantwortung übernehmen_VPA.na:NP.Akk
Weg einschlagen_VPA.na:NP.Akk
Weg gehen_VPA.na:NP.Akk

Qualitative Evaluation: Observations (2/3)

Expected Noun/Verb-Collocations (Examples sorted by verb)

$annehmen\_VPA.na{:}NP.Akk\ Angebot$
$annehmen\_VPA.na{:}NP.Akk\ Ausmass$
$annehmen\_VPA.na{:}NP.Akk\ Form$
$annehmen\_VPA.nd{:}NP.Dat\ Gestalt$
$abgeben\_VPA.na{:}NP.Akk\ Erklärung$
$abgeben\_VPA.na{:}NP.Akk\ Schuss$
$abgeben\_VPA.na{:}NP.Akk\ Stellungnahme$
$abgeben\_VPA.na{:}NP.Akk\ Stimme$

Qualitative Evaluation: Observations (3/3)

Combinations with quantifiers

nichts anhaben_VPA.nad:NP.Akk
nichts bringen_VPA.na:NP.Akk
nichts nützen_VPA.na:NP.Akk

$\Rightarrow$ Statistically significant combinations
$\Rightarrow$ Not in line with the standard collocation definition

Qualitative Evaluation: Problems

Parsing problems

- Collocations with PPs

- *erblassen_VPA.nd:NP.Dat Neid /vor*

- *geraten_VPA.nd:NP.Dat Kontrolle / ausser*

- *zurückbleiben_VPA.nd:NP.Dat Erwartung /hinter*

- Subject/Verb misrecognized

- *fliegen_VPA.na:NP.Akk Fetzen*

- *platzen_VPA.nad:NP.Akk Kragen*

Trivial combinations

- *betragen_VPA.na:NP.Akk Prozent*

- *brauchen_VPA.na:NP.Akk Zeit | Geld | Hilfe*

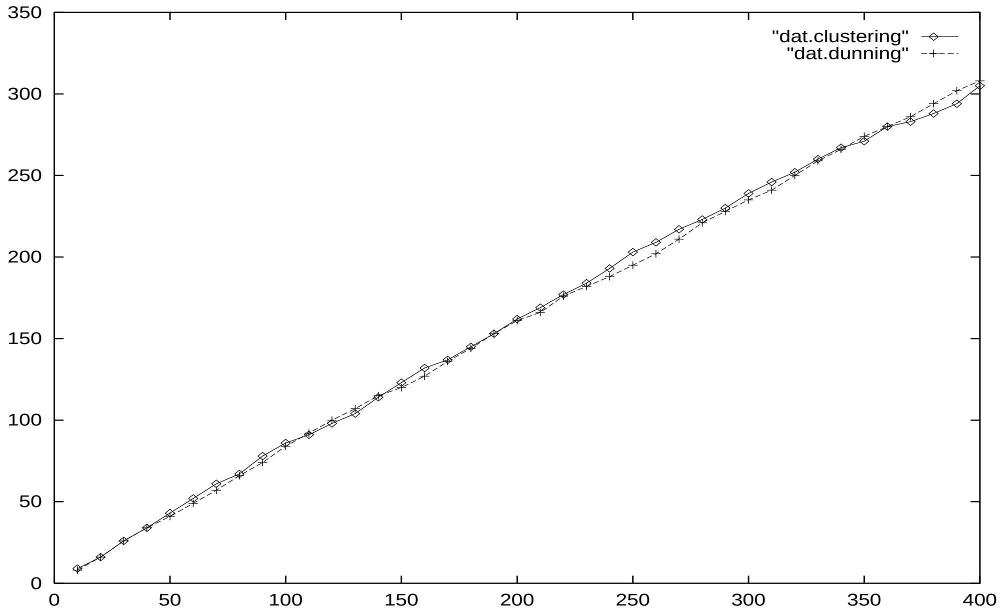
- *bringen_VPA.na:NP.Akk Entlastung | Erfolg | Ergebnis | Lösung*

- *darstellen_VPA.nap:NP.Akk Gefahr*

- *verspinnen_VPA.nd:NP.Dat Schafwolle*

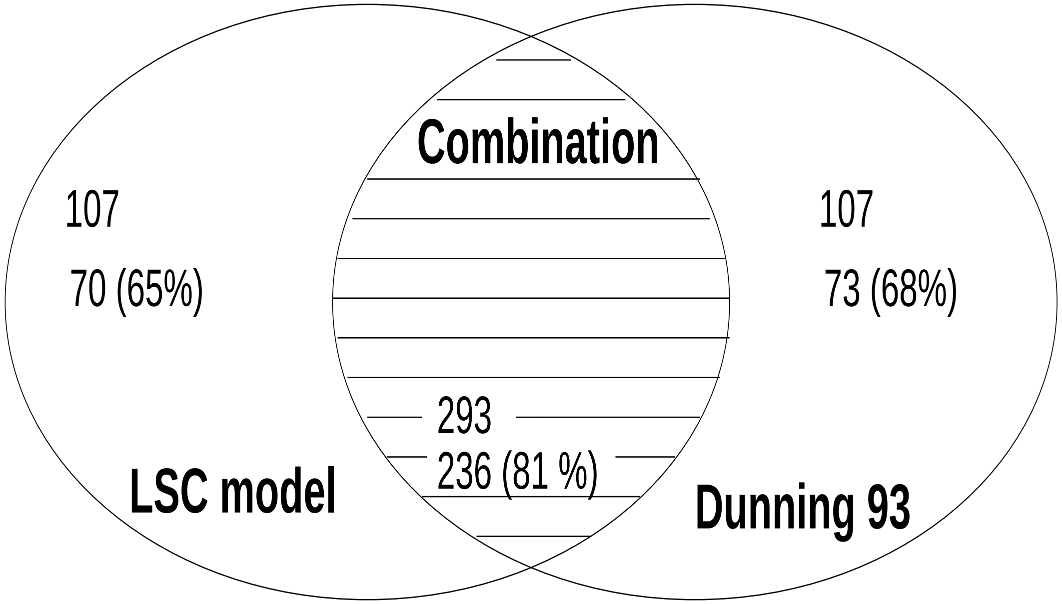
Comparison with Dunning's Scores

- Parsing of 4.1 M word newspaper corpus
- Extract top-400 candidates selected by Dunning's scores and probabilistic clustering
- Merge results from both measures and remove source data
- Select "good" candidates



Boosting performance

- Based on the top 400 candidates:
both systems find ca. 77 % of the "good" candidates individually
- Combination covers 81 % of the "good" candidates



Conclusion

Clustering model:

- provides subcat frames + selectional classes
- allows in addition to identify collocations
- performs on collocations similarly to Dunning's scores

Prerequisites:

- text samples of sufficient size
- robust parser
- But: no need for semantic taxonomy

Next steps:

- more evaluation work is needed to understand specific deficits of individual measures
- more boosting experiments with a combination of two or more measures