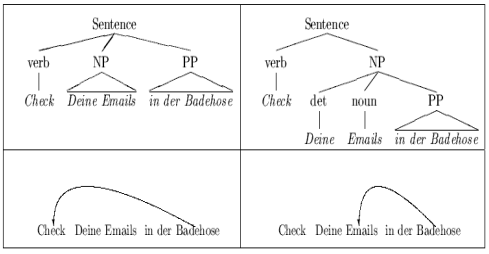
Unsupervised Learning from Corpora

Martin Volk
Universität Zürich
Eurospider Information Technology AG

Overview

1. What is a Prepositional Phrase (PP)?
2. How can we predict PP attachment based on corpus statistics?

PP Attachment Disambiguation



What is a preposition?

- Preposition
  - primary preposition (*in, von, für, mit, auf*)
  - secondary preposition (*gegenüber, trotz, statt*)
- Postposition *(zufolge, ... nach, ... über)*
- Circumposition *(vom ... aus, über ... hinaus)*
- Contracted preposition (*im, zum, zur, ins*)
- Pronominal adverb (*dabei, hiermit, wozu*)
- Reciprocal pronoun (*miteinander, untereinander*)

What is a Prepositional Phrase?

A PP is a sentence constituent

- with a nominal head or
- with a pronominal head (*mit ihm, ohne dessen*) or
- with an adverbial head (*ab hier, von dort*) or
- with an adjectival head (*auf deutsch, für gut*).

Functions of a PP

- Prepositional Object
  - *Formularleser sorgen **für wirtschaftlicheren Zahlungsverkehr***
- Attribute to a noun
  - *ComputerLand soll im Markt ein Synonym **für Dienstleistungen** werden.*
- Attribute to an adjective
  - *Sie sind also **mit dem Geschäftsjahr 1992** zufrieden.*
- Adverbial
  - *Die Kurse schwankten **im Januar** zwischen ...*

Special PPs

- with conjoined prepositions
  - *egal ob **mit oder ohne zusätzliche Hardware***
- with two prepositions
  - ***Innerhalb von 15 Arbeitstagen** hat die Großdruckerei Drescher*
- within idioms
  - *AT&T schlägt bei dem Deal zwei Fliegen **mit einer Klappe**.*

Special PPs

- frozen PPs
  - *und **mit Hilfe** von Unterdruck Herr zu werden versuchen.*
- NP/PPs with noun reduplication
  - *daß er **Schritt für Schritt** in die gewünschte Datenbank einführt.*
- within support verb units
  - *Vorläufig **kommt** ein Investment nur für risikofreudigere Anleger **in Frage**.*

PP attachment ambiguities

- Def.: In German a PP is in an ambiguous position if it follows immediately after a noun in the Mittelfeld.
- We concentrate on verb vs. noun attachment ambiguities. I.e. we disregard adjective attachments.
- We ignore the difference between adjunct vs. complement function.

Systematically ambiguous PPs

- Locative
  - *Mitjubeln beim Champions League Final **in Manchester**?*
- Benefactive
  - *Das Ministerium hat drei Frequenzen **für den Kurzstreckenfunk** freigegeben.*

Previous approaches to PP attachment

- Structural constraints:
  - Minimal attachment: Use as few nonterminals as possible.
  - Right Association: Attach to the most recent phrase.
- Linguistic constraints:
  - Use subcategorisation information (to ask *for*).
  - Use semantic type: temporal PP is attached to the verb.

Statistical approaches

- Supervised
  - Learn attachment preferences from treebank
    For English: up to 84% accuracy
  - Learn from treebank and use WordNet
    For English: up to 88% accuracy
- Unsupervised
  - Learn attachment preferences from shallow parsed corpus
    For English: 80-84% accuracy

Our approach

- Unsupervised statistical approach
  (combined with some linguistics)
- For German
- Learn attachment preferences from shallow parsed corpus
- Use simple cooccurrence measure
  cooc(noun,prep) =
  freq(noun,prep) / freq(noun)

Example of cooccurrence measure

For: ***Check deine Emails in der Badehose***

freq(*Emails, in*) = 50

freq(*Emails*) = 10'000

→cooc(*Emails, in*) = 0.005

freq(*check, in*) = 15

freq(*check*) = 1'000

→cooc(*check, in*) = 0.015

What counts as Cooccurrence?

- For nouns: noun followed by preposition.
- For verbs: verb in the same clause as preposition.
  Requires a clause boundary detector.

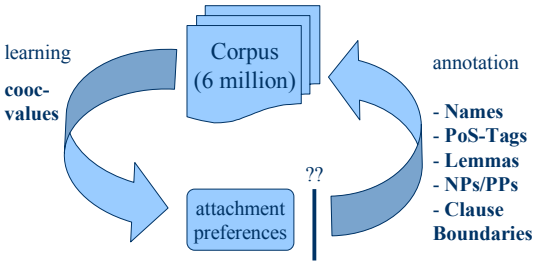
Training Corpus

Annotate a 6 million words computer journal corpus (raw text) through

1. Proper name recognition
2. PoS-Tagging
3. Lemmatisation
4. NP/PP chunking
5. Clause boundary detection

→ Learn cooc(noun,prep) and cooc(verb,prep)

Shallow parsed Corpus



The Computer Zeitung (CZ) treebank

- 3'000 **manually** annotated sentences that contain ambiguous PPs
- German domain specific newspaper texts

→ 4562 PPs in ambiguous positions
1761 with verb attachment (39%)
2801 with noun attachment (61%)

The NEGRA treebank

- 10'000 manually annotated sentences
- German newspaper texts

→ 6064 PPs in ambiguous positions

2664 with verb attachment (44%)

3400 with noun attachment (56%)

Extraction of 5-tuples from treebank sentences

Sentence:

Check (deine Emails) (in der Badehose)

1. Verb: *check*
2. Reference noun N1:
3. Preposition:
4. PP-noun N2:
5. Function:

Extraction of 5-tuples from treebank sentences

Sentence:

Check (deine Emails) (in der Badehose)

1. Verb: *check*
2. Reference noun N1: *Emails*
3. Preposition:
4. PP-noun N2:
5. Function:

Extraction of 5-tuples from treebank sentences

Sentence:

Check (deine Emails) (in der Badehose)

1. Verb: *check*
2. Reference noun N1: *Emails*
3. Preposition: *in*
4. PP-noun N2:
5. Function:

Extraction of 5-tuples from treebank sentences

Sentence:

Check (deine Emails) (in der Badehose)

1. Verb: *check*
2. Reference noun N1: *Emails*
3. Preposition: *in*
4. PP-noun N2: *Badehose*
5. Function:

Extraction of 5-tuples from treebank sentences

Sentence:

Check (deine Emails) (in der Badehose)

1. Verb: *check*
2. Reference noun N1: *Emails*
3. Preposition: *in*
4. PP-noun N2: *Badehose*
5. Function: verb attachment

Experiment setup

1. Learn cooccurrence values from the training corpus
2. Evaluate against the CZ test set
3. Evaluate against NEGRA test set
4. Exchange the training corpus
   1. NZZ as training corpus
   2. Web as training corpus

The learner

- **if N_P sequence**
  then: freq(N)++, freq(N,P)++
  elsif N_x with x <> P
  then: freq(N)++
- **if V...P_1...P_2 in clause**
  then: freq(V)+=2,
  freq(V,P_1)++, freq(V,P_2)++
  elsif V without P in clause
  then: freq(V)++

What is learned?

word W	prep. P	freq(W,P)	freq(W)	cooc(W,P)
Höchstmass	an	13	13	1.000
Hinblick	auf	133	135	0.985
Verweis	auf	21	22	0.955
Umgang	mit	293	307	0.954
logiert	unter	55	56	0.982
paktiert	mit	13	14	0.928
verlautet	aus	16	19	0.842

Disambiguation Algorithm 1

```
if (cooc(N1,P) && cooc(V,P)) then

  if (cooc(N1,P) > cooc(V,P)) then
    noun attachment
  else
    verb attachment
```


Disambiguation Results 1

	correct	incorrect	accuracy
noun att.	925	60	93.91%
verb att.	743	608	55.00%
total	**1668**	**668**	**71.40%**
	2336 / 4143 (57%)		

The noun factor

- Observation: V+P cooccurrence values are too strong
- Needed: a factor to strengthen N+P values
- Based on the overall attachment tendency:
  - cooc(all_Ns, all_Ps) = 0.182
  - cooc(all_Vs, all_Ps) = 0.774

→ noun factor = 0.774 / 0.182 = 4.25

Disambiguation Algorithm 2

if (cooc(N1,P) && cooc(V,P)) then
  if ((cooc(N1,P) * **noun_factor**) > cooc(V,P)) then
    noun attachment
  else
    verb attachment

Disambiguation Results 2

with noun factor 4.25

	correct	incorrect	**accuracy**
noun att.	1377	280	83.10%
verb att.	524	157	76.94%
total	**1901**	**437**	**81.31%**
coverage	**2336 / 4143 (57%)**		

Coverage increase

- Using lemmas
  - verb lemmas
  - noun lemmas (incl. reduction of compounds)
    - *Forschungsinstituts → Institut*

    → coverage of 83% (a gain of 16%)

    → accuracy of 78.13% (a loss of 3%!!)
- Using proper name classes

  → coverage of 86%

  → accuracy of 78.31%

Coverage increase

- not explored
  - diminuitive forms (*Kästchen → Kasten*)
  - different nominalizations (*das Zusammenschalten, die Zusammenschaltung*)
  - number words (*Hundert, Million, Milliarde*)
  - weak nominal prefixes (*Vizepräsident → Präsident*)

Coverage increase

- use partial information
  - evaluate cooc(N,P) against a threshold if cooc(V,P) does not exist (and vice versa)
  - threshold(N) is the average of all cooc(N,P)
- if (cooc(N,P) > threshold(N)) then noun_attachment

Accuracy increase

- Distinguish sure and possible attachment in the learning phase!
- Annotate sure verb attachment (all PPs not following a noun)
  - ***An EU-externe Länder*** *dürfen Daten nur exportiert werden, ...*
  - *Es muß noch* ***vom EU-Ministerrat*** *verabschiedet werden.*

Accuracy increase

- Annotate sure noun attachment (e.g. PPs in copula sentences)
  - *Hintergrund ist die gedämpfte Gewinnerwartung **für 1995**.*
  - *Die Abkehr **von den proprietären Produkten** erzeugt mehr Wettbewerb ...*
- *New Counting:*
  - *sure attach N1: count only for freq(N1,P)*
  - *possible attach: count half for each freq(N1,P) and freq(V,P)*

Integrating linguistic resources

465 support verb units with PPs (B. Krenn)
- *am Anfang stehen*
- *auf Distanz gehen/halten*
- *in Ordnung bringen*

- to annotate sure **verb** attachments in the training corpus
- to decide 97 test cases directly

Integrating linguistic resources

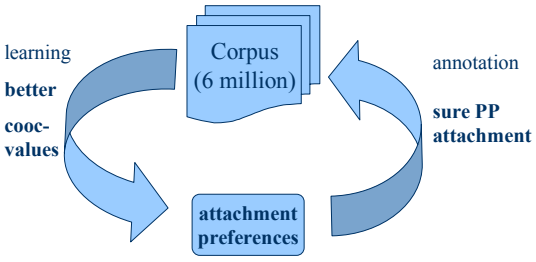
- list of 82 frozen PPs
  (Schröder: Lexikon dt. Präpositionen)
  - *mit Hilfe von*
  - *im Unterschied zu*
  - *aus Mangel an*
- to annotate sure **noun** attachments in the training corpus

➔ only slight improvement in accuracy

Using linguistic regularities

- transfer of cooccurrence value from verb to noun
  - *beteiligen an → Beteiligung an*
  - only for unseen nouns
- distinction between reflexive and regular verb readings
  - *sorgen für*
  - *sich sorgen um*
  - 340 test cases with reflexive verbs
  - no improvement in accuracy

Shallow parsed Corpus



Using pair and triple frequencies

- Hypothesis: Accuracy increases if the PP noun is used.
  - Peter **saw** the **thief with** his own **eyes**.
  - Peter **saw** the **thief with** the red **coat**.

→ *cooc(N1, P, N2) = freq(N1, P, N2) / freq(N1)*

What is learned?

word W	prep. P	noun N2	freq(W,P,N2)	freq(W)	cooc(W,P,N2)
Spreu	von	Weizen	7	16	0.437
Klinke	in	Hand	4.5	11	0.409
Sitz	in	<loc>	134.5	418	0.323
Nachfolge	von	<person>	27.5	86	0.319
Made	in	Germany	7.5	24	0.312

Disambiguation Algorithm 3

if (cooc(N1,P,N2) && cooc(V,P,N2)) then
  if (cooc(N1,P,N2) * noun_factor) > cooc(V,P,N2)
    then: noun attachment
    else: verb attachment
elsif (cooc(N1,P) && cooc(V,P)) then
  if ((cooc(N1,P) * noun_factor) > cooc(V,P))
    then: noun attachment
    else: verb attachment

Results for unsupervised method

decision level	# of cases	accuracy
support verb unit	97	100.00%
triple comparison	953	84.36%
pair comparison	2813	79.95%
cooc(N1,P) > thr.	74	85.13%
cooc(V,P) > thr.	91	84.61%
total	**4028 (coverage: 90%)**	**81.67%**

Using GermaNet

- using synonym classes to cluster nouns
- if noun belongs to more than one class, count it for all of them
- in disambiguation use highest cooccurrence value
  - cooc(*Kunde, über*) = 0.00374
  - cooc(*Kunde_Wissen_Kenntnis, über*) = 0.00909
  - cooc(*Kunde_Kundin, über*) = 0.00374
- impact of GermaNet usage is not visible in the attachment results

Pronominal adverbs

- are placeholders for (complement) PPs
  - *Sie warten auf solche Geräte / darauf ...*
  - *Sie warten auf die Informatiker / *darauf ...*
  - *Sie spekulieren auf dem Parkett / *darauf über ...*
- can be ambiguous but
  - in CZ test set: 83% verb attachments (of 43)
  - in NEGRA test set: 80% verb attachments (of 109)

→ attach to noun only with strong evidence

Comparative phrases

- are a borderline case of PPs
  - *... befasst sich mit der Sprache **als Steuermedium** für PCs.*
  - *... erweist sich die CD-ROM **als flexibles Medium**.*
  - *... könnten die Daten **wie herkömmliche Informationsquellen** verwenden.*
- Tendency:
  - in CZ test set: 66% verb attachments (of 48)
  - in NEGRA test set: 75% verb attachments (of 145)