Supervised Learning from Corpora

Martin Volk
Universität Zürich

Overview

1. Transformation-Based Learning
2. The Back-Off Method
3. A combination of supervised and unsupervised results

Training material

- the NEGRA test set (6064 test cases)
- the CZ test set (4562 test cases)

➔ as basis for small training sets!!

The sparse data problem

Many quadruples will occur rarely!

Therefore: clustering is needed

- verbs → lemmas
- contracted prepositions → base forms
- proper names → class labels
- numbers → number tag
- nouns → lemmas (of last compound element)

Transformation-Based Learning

- developed by Eric Brill for Part-of-Speech Tagging
- idea: learn transformation rules from the manually disambiguated cases based on rule templates

Transformation-Based Learning for PP-attachment

- idea: learn transformation rules from the manually disambiguated cases based on rule templates
  - start with "noun attachment"
  - in each step determine the rule that contributes most to the correction of the training set
  - the rule templates use the quadruple (V, N1, P, N2) and can access
    - one specific word (4 templates) or
    - any combination of two words (6 templates) or
    - any triple that contains the preposition (3 templates)

Examples for learned rules

- change from noun att. to verb att. if N1 = <person>
- change from noun att. to verb att. if P = *auf*
- change from verb att. to noun att. if N1 = <Person> && P = *von*

- Rules learned from CZ test set and NEGRA test set are very similar!
- BUT: Rule learning even for our small training corpus takes hours!!

Application of the Rules

- Start with the default attachment decision and correct with the transformation rules.

Disambiguation Results for the Transformation-Based Method

Training corpus	Test set	Accuracy
NEGRA	CZ	**72.3%**
NEGRA + 4/5 CZ	1/5 CZ * 5	**∅ = 76.4%**

Back-off Method

- by Collins and Brooks
- idea: learn attachment tendencies from manually disambiguated cases
- in case of missing information back-off to the next level
  - quadruples
  - triples (which include the preposition)
  - pairs (which include the preposition)
  - prepositions alone

Back-off Algorithm

if (freq(V,N1,P,N2) > 0) then
  if (freq(noun_att, V,N1,P,N2) / (freq(V,N1,P,N2)) > 0.5
    then noun attachment
    else verb attachment
elsif ((freq(V,N1,P) + freq(V,P,N2) + freq(N1,P,N2)) > 0) then
  if ((freq(noun_att, V,N1,P) + freq(noun_att, V,P,N2) + freq(noun_att, N1,P,N2)) / (freq(V,N1,P) + freq(V,P,N2) + freq(N1,P,N2))) > 0.5
    then noun attachment
    else verb attachment
elsif ((freq(V,P) + freq(P,N2) + freq(N1,P)) > 0) then
  ...

Back-off Example

V: geben N1: Firma P: in N2: <GEO> Fct:MNR

freq(n_att, V,N1,P) :	
freq(n_att, V,P,N2) :	1
freq(n_att, N1,P,N2):	
Sum	**1**
freq(V,N1,P) :	
freq(V,P,N2) :	5
freq(N1,P,N2):	
Sum	**5**
Ratio	**0.2**

Results for the Back-off method

decision level	# of cases	accuracy
quadruples	8	100.00%
triples	329	88.75%
pairs	3040	75.66%
preposition	1078	64.66%
default	14	64.29%
total	**4469 (coverage: 100%)**	**73.98%**

Results for the Back-off Method

Training corpus	Test set	Accuracy
NEGRA	CZ	74.0%
CZ	NEGRA	68.3%
NEGRA + 4/5 CZ	1/5 CZ * 5	∅ = 79.4%

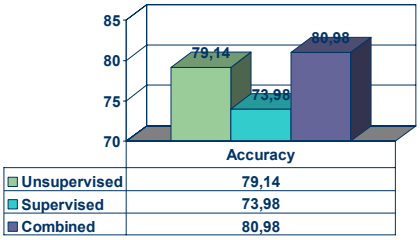
Results for Our Unsupervised method

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%**

Intertwined Combination

Unsupervised	Supervised
1. Support verb unit	1. -
2. -	2. Quadruples
3. -	3. Triples
4. Triple comparison	4. -
5. Pair comparison	5. -
6. Threshold compar.	6. -
7. -	7. Pairs
8. -	8. Preposition
9. -	9. Default

Comparison of results



Conclusions

- Supervised methods lead to better results than unsupervised methods given enough treebank material from the right domain.
- Unsupervised method is as good as supervised method over small training corpus.
- Combination of unsupervised and supervised leads to the best results.

Conclusions

- Beware: V,N1,P,N2 makes the PP attachment task look easier than it is!!
- It reduces the PP attachment task to a simple case.
- But often: sequences of NPs and PPs:
  - V_NP_PP_PP or
  - V_NP_NP_PP