What makes sense?
Searching for strong WSD predictors in Croatian

Nikola Bakarić
Jasmina Njavro
Nikola Ljubešić

Department of Information Sciences
Faculty of Humanities and Social Sciences
Zagreb, Croatia

8 Nov 2007

Ambiguity in language

- natural language ambiguous - problem in automated processing

Ambiguity in language

- natural language ambiguous - problem in automated processing
- ambiguity on more levels - lexical, syntactic, semantic

Ambiguity in language

- natural language ambiguous - problem in automated processing
- ambiguity on more levels - lexical, syntactic, semantic
- one lexeme can have more than one meaning

INFuture 2007
The Future of Information Sciences

Ambiguity in language

- natural language ambiguous - problem in automated processing
- ambiguity on more levels - lexical, syntactic, semantic
- one lexeme can have more than one meaning
  - polysemy - word or phrase with multiple related meanings
  - homonymy - group of words that share the same spelling (or pronunciation)

Ambiguity in language

- natural language ambiguous - problem in automated processing
- ambiguity on more levels - lexical, syntactic, semantic
- one lexeme can have more than one meaning
  - polysemy - word or phrase with multiple related meanings
  - homonymy - group of words that share the same spelling (or pronunciation)
- borders between homonymy and polysemy are often unclear and change over time

Ambiguity in language

- natural language ambiguous - problem in automated processing
- ambiguity on more levels - lexical, syntactic, semantic
- one lexeme can have more than one meaning
  - polysemy - word or phrase with multiple related meanings
  - homonymy - group of words that share the same spelling (or pronunciation)
- borders between homonymy and polysemy are often unclear and change over time
- SENSEVAL - inter-annotator agreement only around 60%

Natural language processing and WSD

- use of word sense disambiguation (WSD) in NLP
  - information retrieval
  - automated indexing
  - machine translation
  - ...

Natural language processing and WSD

- use of word sense disambiguation (WSD) in NLP
  - information retrieval
  - automated indexing
  - machine translation
  - ...
- two major approaches in NLP
  - deterministic approach - rule-based
  - stochastic approach - based on probability

Natural language processing and WSD

- use of word sense disambiguation (WSD) in NLP
  - information retrieval
  - automated indexing
  - machine translation
  - ...
- two major approaches in NLP
  - deterministic approach - rule-based
  - stochastic approach - based on probability
- stochastic approach
  - supervised methods - training a model on annotated data set
  - unsupervised methods - clustering on unannotated data set

Preparing the data

- corpus: Vjesnik daily newspaper, on-line edition
- consists of:
  - 187 323 articles
  - 82 826 497 words

Preparing the data

- corpus: Vjesnik daily newspaper, on-line edition
- consists of:
  - 187 323 articles
  - 82 826 497 words
- two separate lists: “miš” (“mouse” – the first list) and “stanica” (“cell” – the second list)

INFuture 2007
The Future of Information Sciences

Preparing the data

- corpus: Vjesnik daily newspaper, on-line edition
- consists of:
    - 187 323 articles
    - 82 826 497 words
- two separate lists: "miš" ("mouse" – the first list) and "stanica" ("cell" – the second list)
- randomly divided into ten sets - used for 10-fold cross-validation

Preparing the data

- corpus: Vjesnik daily newspaper, on-line edition
- consists of:
  - 187 323 articles
  - 82 826 497 words
- two separate lists: “miš” (“mouse” – the first list) and “stanica” (“cell” – the second list)
- randomly divided into ten sets - used for 10-fold cross-validation
- corpus verticalised, sentence boundaries marked

Manual annotation

- manualy determining word sense ("miš" - 8, "stanica" - 6)

Manual annotation

- manualy determining word sense ("miš" - 8, "stanica" - 6)
- manual annotation of 1000 occurences for each lexeme
  60% by both annotators, the rest separatley
- due to strong polysemy -inter-annotator agreement was 100%

INFuture 2007
The Future of Information Sciences

Naïve Bayes classifier

- simple probabilistic learning algorithm

Naïve Bayes classifier

- simple probabilistic learning algorithm
- calculates the a priori and the a posteriori conditional probability of an event in the training corpus, decision by MAP (maximum a posteriori) rule

$$k(x_1, x_2, ...x_P) = \underset{y}{\operatorname{argmax}} P(Y = y) \prod_{p=1}^{P} p(X_p = x_p | Y = y)$$

Naïve Bayes classifier

- simple probabilistic learning algorithm
- calculates the a priori and the a posteriori conditional probability of an event in the training corpus, decision by MAP (maximum a posteriori) rule

$$k(x_1, x_2, ...x_P) = \underset{y}{\operatorname{argmax}} P(Y = y) \prod_{p=1}^{P} p(X_p = x_p | Y = y)$$

- no feature selection – all types are features

Naïve Bayes classifier

- simple probabilistic learning algorithm
- calculates the a priori and the a posteriori conditional probability of an event in the training corpus, decision by MAP (maximum a posteriori) rule

$$k(x_1, x_2, ...x_P) = \operatorname*{argmax}_y P(Y = y) \prod_{p=1}^{P} p(X_p = x_p | Y = y)$$

- no feature selection – all types are features
- disadvantage: assumes variables are independent events

INFuture 2007 The Future of Information Sciences

Naïve Bayes classifier

- simple probabilistic learning algorithm
- calculates the a priori and the a posteriori conditional probability of an event in the training corpus, decision by MAP (maximum a posteriori) rule

$$k(x_1, x_2, ...x_P) = \underset{y}{\operatorname{argmax}}\, P(Y = y) \prod_{p=1}^{P} p(X_p = x_p | Y = y)$$

- no feature selection – all types are features
- disadvantage: assumes variables are independent events
- advantage: not affected by the curse of dimensionality, produces good results without feature selection

Naïve Bayes classifier

- simple probabilistic learning algorithm
- calculates the a priori and the a posteriori conditional probability of an event in the training corpus, decision by MAP (maximum a posteriori) rule

$$k(x_1, x_2, \ldots x_P) = \underset{y}{\operatorname{argmax}}\, P(Y = y) \prod_{p=1}^{P} p(X_p = x_p | Y = y)$$

- no feature selection – all types are features
- disadvantage: assumes variables are independent events
- advantage: not affected by the curse of dimensionality, produces good results without feature selection
- we do not observe the absoulute accuracy, but the relative shift in regards to the environment size

Experiment

- goal: determine the effect of lexeme's sorroundings on its meaning

Experiment

- goal: determine the effect of lexeme's sorroundings on its meaning
- we observed the following in regards of accuracy:
  - changing window size - including 1-50 words left and right

Experiment

- goal: determine the effect of lexeme's sorroundings on its meaning
- we observed the following in regards of accuracy:
  - changing window size - including 1-50 words left and right
  - changing window distance - including 1 word left and right while changing distance

- goal: determine the effect of lexeme's sorroundings on its meaning
- we observed the following in regards of accuracy:
  - changing window size - including 1-50 words left and right
  - changing window distance - including 1 word left and right while changing distance
  - evaluating the accuracy of one-sense-per-discourse method

Experiment

- goal: determine the effect of lexeme's sorroundings on its meaning
- we observed the following in regards of accuracy:
  - changing window size - including 1-50 words left and right
  - changing window distance - including 1 word left and right while changing distance
  - evaluating the accuracy of one-sense-per-discourse method
  - evaluating impact of sentence border on determining lexeme's sense

INFuture 2007
The Future of Information Sciences

Experiment

- goal: determine the effect of lexeme's sorroundings on its meaning
- we observed the following in regards of accuracy:
  - changing window size - including 1-50 words left and right
  - changing window distance - including 1 word left and right while changing distance
  - evaluating the accuracy of one-sense-per-discourse method
  - evaluating impact of sentence border on determining lexeme's sense
- evaluation by accuracy and standard error through 10-fold cross-validation

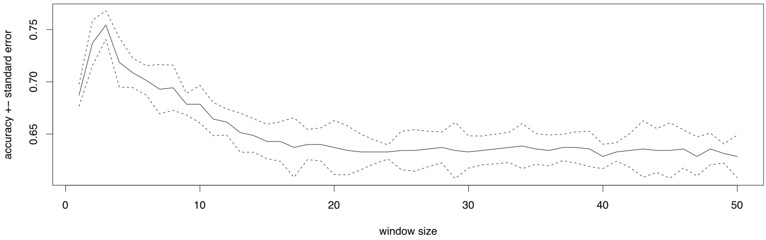
$$accuracy = \frac{a+d}{a+b+c+d}$$

$$S_E = \frac{\sigma}{\sqrt{N}}$$

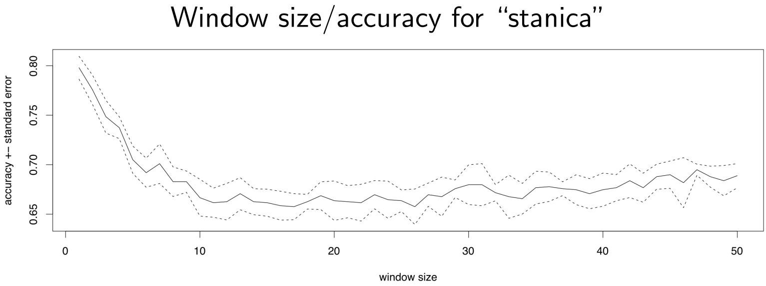
INFuture 2007 The Future of Information Sciences

Results

Window size/accuracy for "miš"

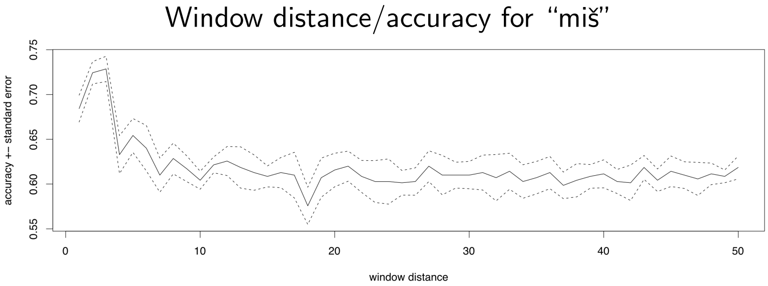


Results


Window size/accuracy for "stanica"
accuracy +− standard error
0.65
0.70
0.75
0.80
0
10
20
30
40
50
window size


INFuture 2007
The Future of Information Sciences

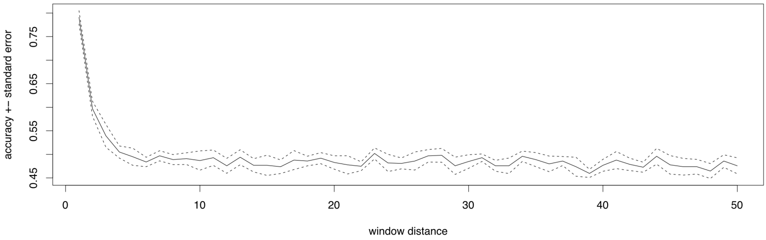
Results



Window distance/accuracy for “miš”
accuracy +- standard error
0.55
0.60
0.65
0.70
0.75
0
10
20
30
40
50
window distance

Results

Window distance/accuracy for “stanica”



Results

One-sense-per-discourse method

	Applicability	Accuracy
“miš”	28.92%	88.98%
“stanica”	26.31%	97.10%

Results

Accuracy with standard error in relation to 3 tokens before/after observed lexeme sentence boundary

	Before sentence boundary	After sentence boundary
“miš”	68,00%±1,52%	64,14%±2,09%
“stanica”	57,37%±1,75%	57,27%±1,18%

Conclusion

- strong predictors in window size 1-5
  - both window size and window distance experiments confirm the immediate sorrounding of the lexeme as the most informative when determining strong WSD predictors

Conclusion

- strong predictors in window size 1-5
  - both window size and window distance experiments confirm the immediate sorrounding of the lexeme as the most informative when determining strong WSD predictors
- good results when applying one-sense-per-discourse method
  - when the observed lexeme appears more than once in a discourse, the probability is very high that it will have the same sense

Conclusion

- strong predictors in window size 1-5
  - both window size and window distance experiments confirm the immediate sorrounding of the lexeme as the most informative when determining strong WSD predictors
- good results when applying one-sense-per-discourse method
  - when the observed lexeme appears more than once in a discourse, the probability is very high that it will have the same sense
- sentence border does not appear to be significant for strong WSD predictors

contact:
nbakaric@ffzg.hr
jnjavro@ffzg.hr
nljubesi@ffzg.hr

Questions?