Optimization in Machine Learning of Word Sense Disambiguation

Walter Daelemans

daelem@uia.ua.ac.be

http://cnts.uia.ac.be

CNTS, University of Antwerp

ILK, Tilburg University

Meaning-03, April 2003

Work in progress with

Véronique Hoste, Fien De Meulder
(CNTS, Antwerp)

Bart Naudts

(Computer Science, Antwerp)

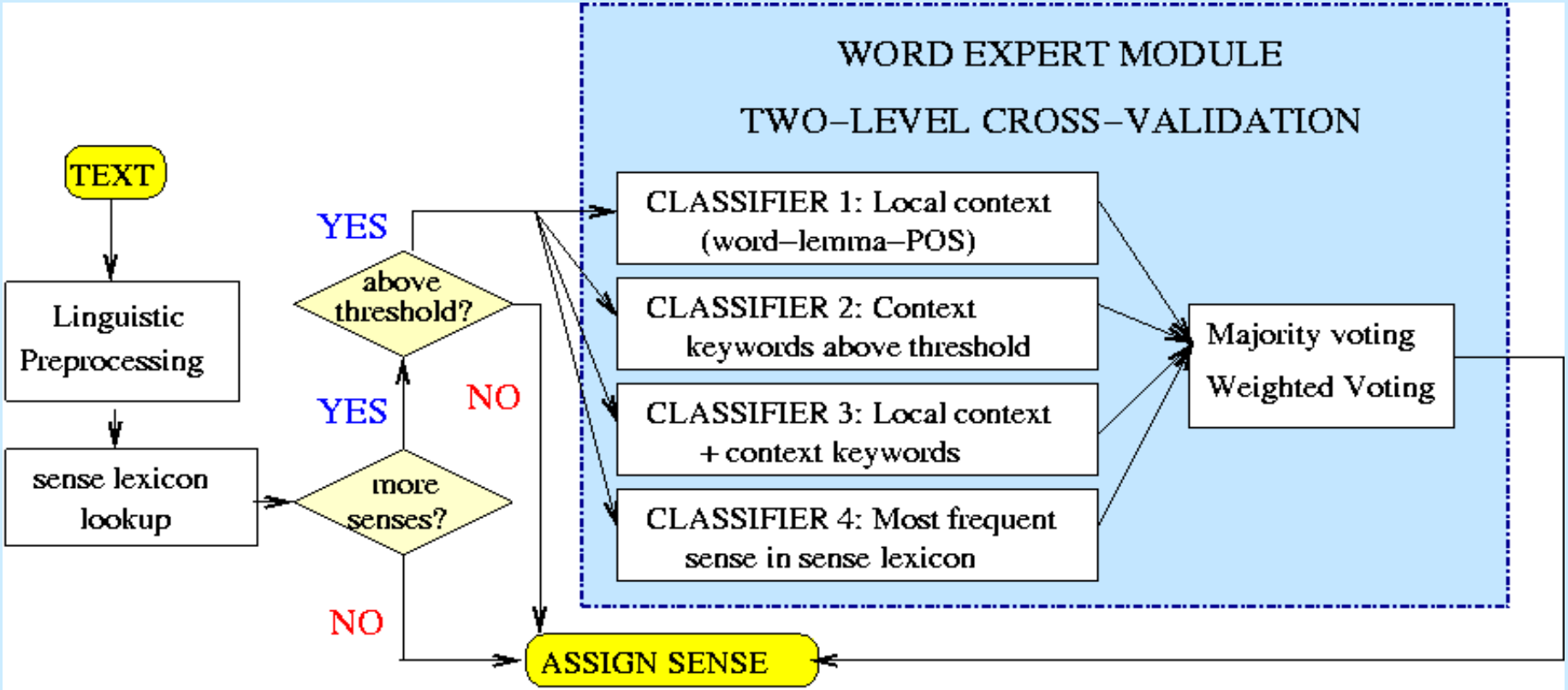
Outline

- Tilburg-Antwerp learning word expert approach to WSD
- Effect of feature selection and algorithm parameter optimization on WSD accuracy
- The larger problem of comparative machine learning experiments
- Using Genetic Algorithms for optimization
- Conjectures: where to invest effort for ML of WSD (and NLP in general)?

The Meaning project

- Great:
  - Advanced ML technology applied to the tasks
  - The Knowledge Acquisition / WSD / text analysis tools interaction
  - Productivity of the project members
- But:
  - Sense inventories are task and domain-dependent
  - Reliability of comparative machine learning experiments is debatable (this presentation)

CNTS-ILK approach all-words task



Information Sources

- Local information: 3 word forms to left and right + POS + (lemma), e.g.

> no_matter RB whether IN he PRP has have VBZ short JJ or long JJ have%2:42:00

- Keyword information: disambiguating keywords in a context of three sentences. (Ng and Lee, 1996)

> A word is a keyword for a given sense, if
> 1. the word occurs more than a predefined minimum number of times with that sense
> 2. $p(s|k) \geq$ predefined minimum probability

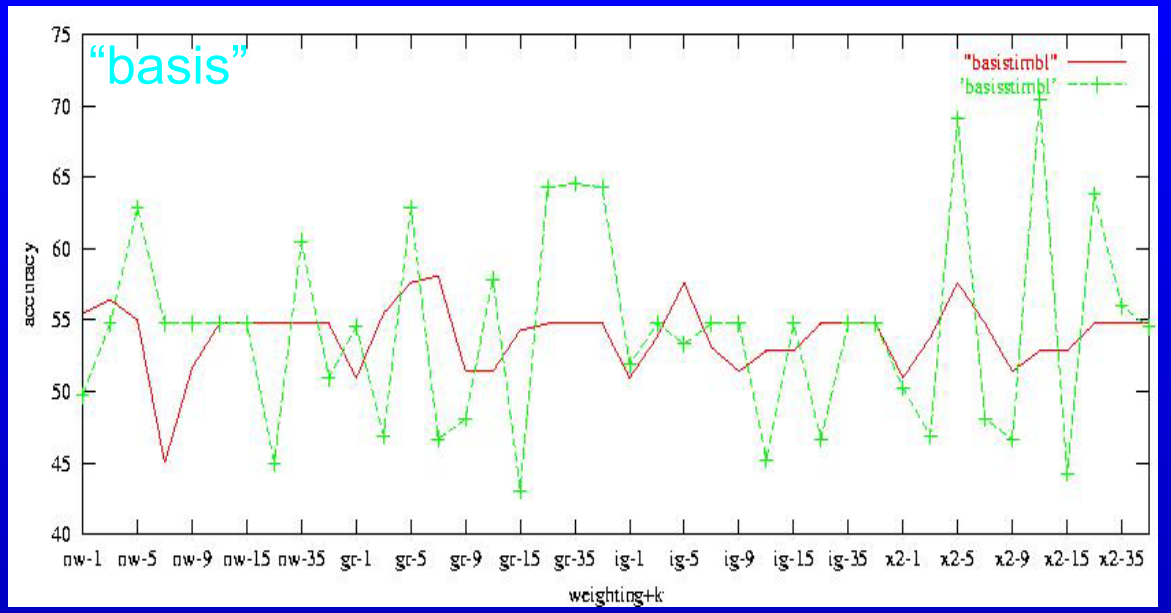
POS versus Information Source

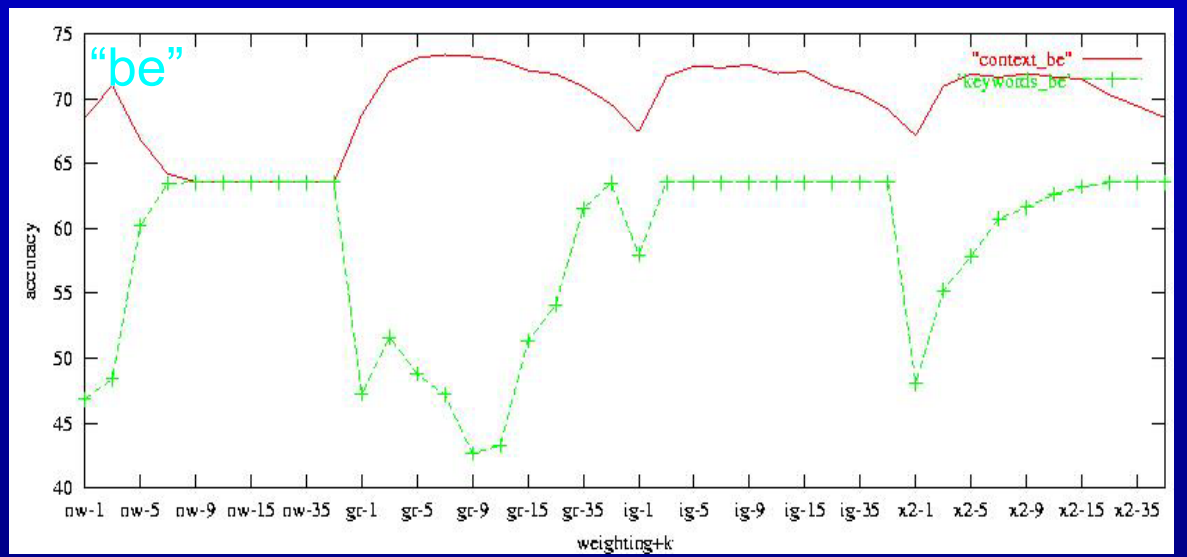
POS	Basel.	Local cont.	Keyw.	Local cont. + keyw.	Maj. voting	Maj. voting (no def.)	Weigh. voting	Weigh. voting (no def.)
NN	64.2	71.4	**74.2**	69.3	69.3	72.7	73.4	73.8
VB	56.9	64.3	63.8	60.1	60.8	63.6	**64.6**	**64.6**
JJ	66.3	72.2	**73.8**	70.4	70.4	72.8	73.3	73.6
RB	70.0	**76.6**	74.5	73.1	72.5	74.9	75.5	75.4
All	61.7	70.1	70.0	66.9	66.5	69.9	69.9	**70.3**

Optimization of algorithm parameters per WE

- Optimizing algorithm parameters for each expert independently in senseval-1 lexical sample accounted for an average 14.4% accuracy increase compared to same settings for all experts
  - Veenstra et al. 2000 (CHUM)
- Optimizing algorithm parameters in interaction with selected features (partially controlled for in senseval-2 all words), accounts for estimated additional accuracy increase greater than 3%
  - Hoste et al. 2002 (NLE)

Influence of the choice of information source on the accuracy for different feature weighting methods and k values.

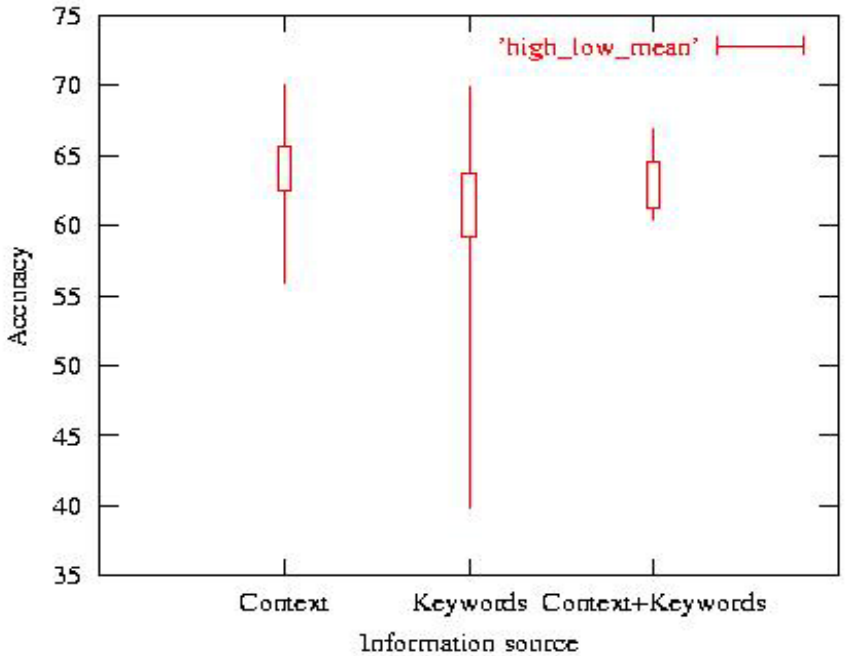




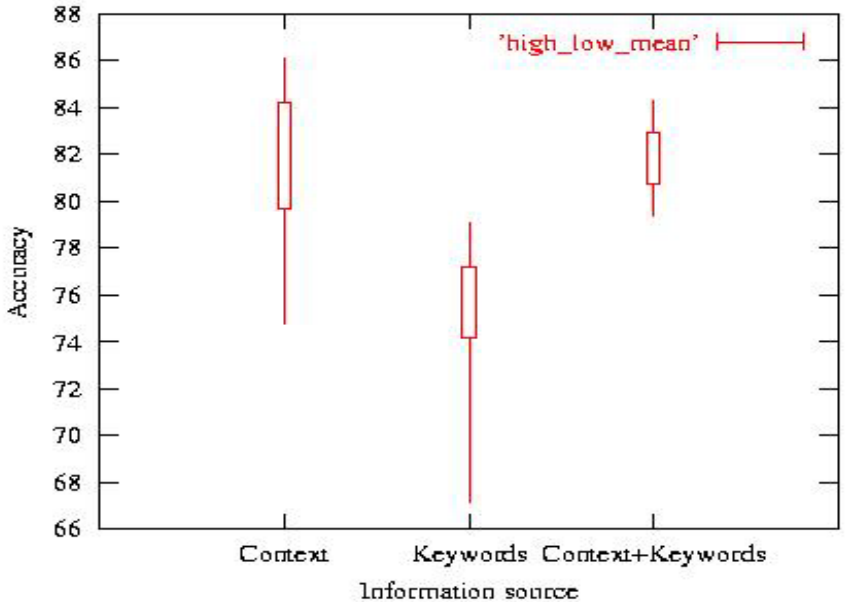
Optimal parameter settings for one WE cannot be generalized to other WE

English

Results of the three MBL classifiers over all parameter settings over all word-experts (weighted by frequency)



No overall optimal
- information source
- parameter setting



Dutch

Conclusion

Changing any of the architectural variables can lead to **large fluctuations** in the generalization accuracy

Cross-validating algorithm parameters and information sources should be included as a first step in constructing WSD systems, and NLP systems in general

But it’s even worse …

What are the goals of Machine Learning in NLP?

- Machine Learning may alleviate the problems of mainstream statistical methods in NLP
- Which method has the right "bias" for NLP?
- From which information sources do the best ML methods benefit most?
- *A priori*, nothing can be said about this (Hume's problem of induction)
- These questions have to be solved empirically

Result: focus on Comparative ML experiments in NLP

- Evaluate bias of ML method for some (class of) NLP tasks (e.g. WSD)
- Evaluate the role of different information sources in solving a ML of NL task (e.g. WSD)
- Examples:
  - EMNLP, CoNLL, ACL, …
  - Competitions:
    - SENSEVAL
    - CoNLL shared tasks
    - TREC / MUC / DUC / …

What influences the outcome of a (comparative) ML experiment?

- Information sources
  - feature selection
  - feature representation (data transforms)
- Algorithm parameters
- Training data
  - sample selection
  - sample size (Banko & Brill)
- Combination methods
  - bagging, boosting
  - output coding

- Interactions
  - Algorithm parameters and sample selection
  - Algorithm parameters and feature representation
  - Feature representation and sample selection
  - Sample size and feature selection
  - Feature selection and algorithm parameters
  - …

Current Practice Comparative ML Experiments

- Methodology: k-fold cross-validation, McNemar, paired t-test, learning curves, etc.
- Use default algorithm parameters
- Sometimes: algorithm parameter optimization
- Sometimes: feature selection
- Rarely: first feature selection then parameter optimization
- Never: interleaved feature selection and parameter optimization

= combinatorial optimization problem

Hypotheses

The observed difference in accuracy between two algorithms can be easily dwarfed by accuracy differences resulting from interactions of algorithm parameter settings and feature selection.

The observed direction of difference in accuracy of a single algorithm with two sets of features can easily be reversed by the interaction with algorithm parameter settings

Back to WSD
Comparative research

- *Mooney*, EMNLP-96
  - NB & perceptron > DL > MBL ~ Default
  - "Line", no algorithm parameter optimization, no feature selection, no MBL feature weighting, …
- *Ng*, EMNLP-97
  - MBL > NB
  - No cross-validation
- *Escudero, Marquez, & Rigau*, ECAI-00
  - MBL > NB
  - No feature selection
- *Escudero, Marquez, Rigau*, CoNLL-00
  - LazyBoosting > NB, MBL, SNoW, DL

- *Zavrel, Degroeve, Kool, Daelemans,* TWLT-00
  - Senseval-1
  - SVM > MBL > ME > NB > FAMBL > RIP > WIN > C4.5
- *Lee & Ng,* EMNLP-02
  - State-of-the-art comparative research
  - Studies different knowledge sources and different learning algorithms and their interaction
  - Senseval-1 and senseval-2 data (lexical sample, English)
  - All knowledge sources better than any 1
  - SVM > Adb, NB, DT
  - No algorithm parameter optimization
  - No interleaved feature selection and algorithm parameter optimization
- Meaning deliverable WoP6.8
  - SVM ~ Adb > MBL > NB ~ DL > default

Experiment 1

- Investigate the effect of
  - algorithm parameter optimization
  - feature selection (heuristic forward selection)
  - interleaved feature selection and parameter optimization
- … on the comparison of two inductive algorithms (lazy and eager)
- … for WSD

Algorithms compared

- Ripper
  - *Cohen, 95*
  - Rule Induction
  - Algorithm parameters: different class ordering principles; negative conditions or not; loss ratio values; cover parameter values
- TiMBL
  - *Daelemans/Zavrel/van der Sloot/van den Bosch, 98*
  - Memory-Based Learning
  - Algorithm parameters: ib1, igtree; overlap, mvdm; 5 feature weighting methods; 4 distance weighting methods; 10 values of k

Line (all - sampled) words

	Ripper	TiMBL
Default	63.9 - 40.4	60.2 - 59.1
Optimized parameters	70.2 - 61.2	63.4 - 66.4
Optimized features	63.9 - 40.9	62.7 - 60.3
Optimized parameters + FS	91.3 - 63.3	64.5 - 66.7

Line (all - sampled) words + tags

	Ripper	TiMBL
Default	63.8 - 41.4	57.8 - 56.9
Optimized parameters	71.6 - 60.5	64.3 - 67.3
Optimized features	64.7 - 41.6	62.7 - 61.5
Optimized parameters + FS	76.4 - 61.1	64.9 - 68.1

POS Tagging (known-unknown)

	Ripper	TiMBL
Default	93.1 - 76.1	93.0 - 76.3
Optimized parameters	93.9 - 78.1	95.2 - 82.2
Optimized features	93.3 - 76.3	95.0 - 76.5
Optimized parameters + FS	94.5 - 78.1	96.5 - 82.2

Generalizations?

- Accuracy landscapes are not regular
- In general, best features or best parameter settings are unpredictable for a particular data set and for a particular ML algorithm
- Note: these are heuristic results, exhaustive exploration of the accuracy landscape is computationally not feasible

Experiment 2

- Investigate the effect of
  - algorithm parameter optimization
- … on the comparison of different knowledge sources for one inductive algorithm (TiMBL)
- … for WSD
  - Local context
  - Local context and keywords
  - Local context and pos tags

do

	Local Context	+ keywords
Default	49.0	47.9
Optimized	60.8	61.0

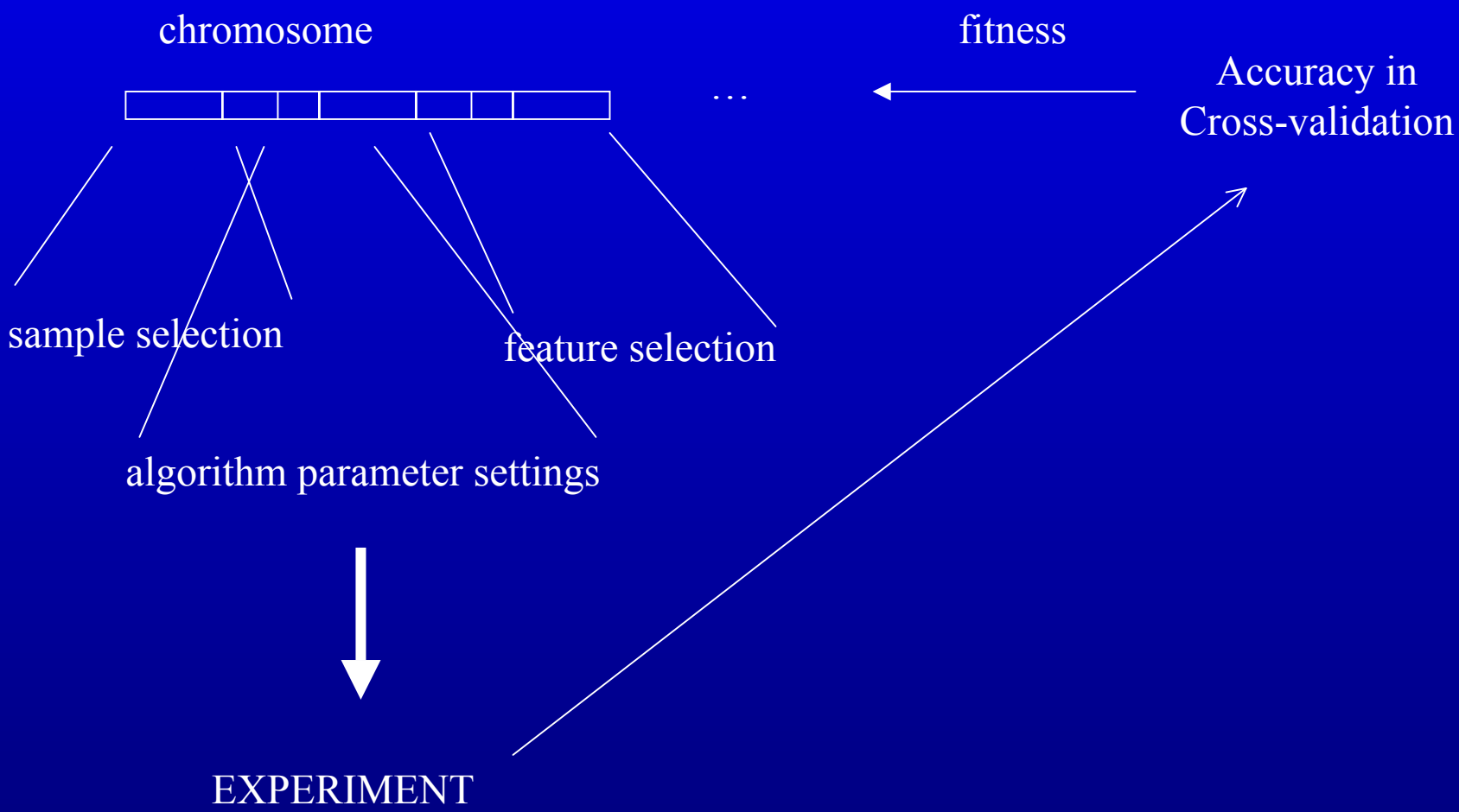
line (all - sampled)

	words	+ pos tags
Default	60.2 - 59.1	57.8 - 56.9
Optimized	64.5 - 66.7	64.9 - 68.1

Interpretation?

- Exhaustive interleaved algorithm parameter optimization and feature selection is in general computationally intractable
- There seem to be no generally useful heuristics to prune the experimental search space
- In addition, there may be interaction with sample selection, sample size, feature representation, etc.
- *Genetic Algorithms* seem to be a good choice in cases like this

Genetic Algorithms



Mapping experiments to GA (TiMBL)

- Each feature represented by one gene
  - Value: selected (1), deselected (0), mvdm (2)
- Weighting metric represented by one gene
- Value of **k** represented by one gene
- Distance weighting method represented by one gene
- Mutation and crossover operators special-purpose
- Complete chromosome maps to experiment
- Accuracy is fitness of chromosome in ten-fold CV
- Chromosomes selected and recombined according to fitness

First Results

- Population Size 100, 20 generations
- Ten-fold cross validation for determining fitness

Word Expert	Default	Best at 1	Best at 20
bar	36.46	43.90	49.47
channel	29.57	34.88	43.00
develop	19.50	28.57	28.57
natural	33.80	37.45	47.87
post	54.93	61.14	65.67

Conclusion

- Optimizing algorithm parameter setting and feature selection interaction has a huge effect on generalization accuracy and on the comparison of ML algorithms and information sources
- Current published results are methodologically correct but nevertheless unreliable
- For many problems and algorithms, this optimization is computationally not feasible
- GAs may be one solution
- Parameterless algorithms ?
- Is the ML of NL field in need of new goals?

Fantasy: where will progress in WSD come from?

All words (~65%)	Senseval-4	Senseval-5
More computing power for optimization	+5%	+10%
More annotated data / better tools	+10%	+20%
Unannotated data	+5%	+10%
Combined	+15%	+25% (Solved!)