Improving Morphosyntactic Tagging of Slovene by Tagger Combination

Jan Rupnik
Miha Grčar
Tomaž Erjavec

Jožef Stefan Institute

Outline

- Introduction
- Motivation
- Tagger combination
- Experiments

POS tagging

Part Of Speech (POS) tagging: assigning morphosyntactic categories to words

Slovenian POS

- multilingual MULTEXT-East specification
- almost 2,000 tags (morphosyntactic descriptions, MSDs) for Slovene
- Tags: positionally coded attributes
- Example: MSD **Agufpa**
  - Category = Adjective
  - Type = general
  - Degree = undefined
  - Gender = feminine
  - Number = plural
  - Case = accusative

State of the art: Two taggers

- Amebis d.o.o. proprietary tagger
  - Based on handcrafted rules
- TnT tagger
  - Based on statistical modelling of sentences and their POS tags.
  - Hidden Markov Model tri-gram tagger
  - Trained on a large corpus of anottated sentences

Statistics: motivation

•Different tagging outcomes of the two taggers on the JOS corpus of 100k words

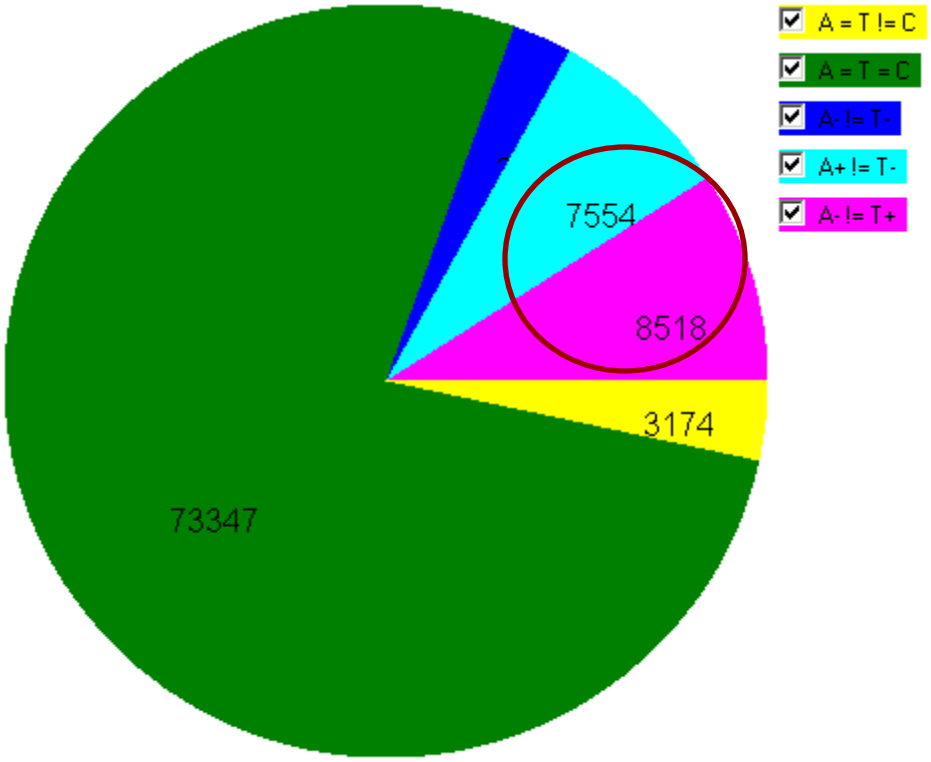
•Green: proportion of words where both taggers were correct

•Yellow: Both predicted the same, incorrect tag

•Blue: Both predicted incorrect but different tags

•Cyan: Amebis correct, TnT incorrect

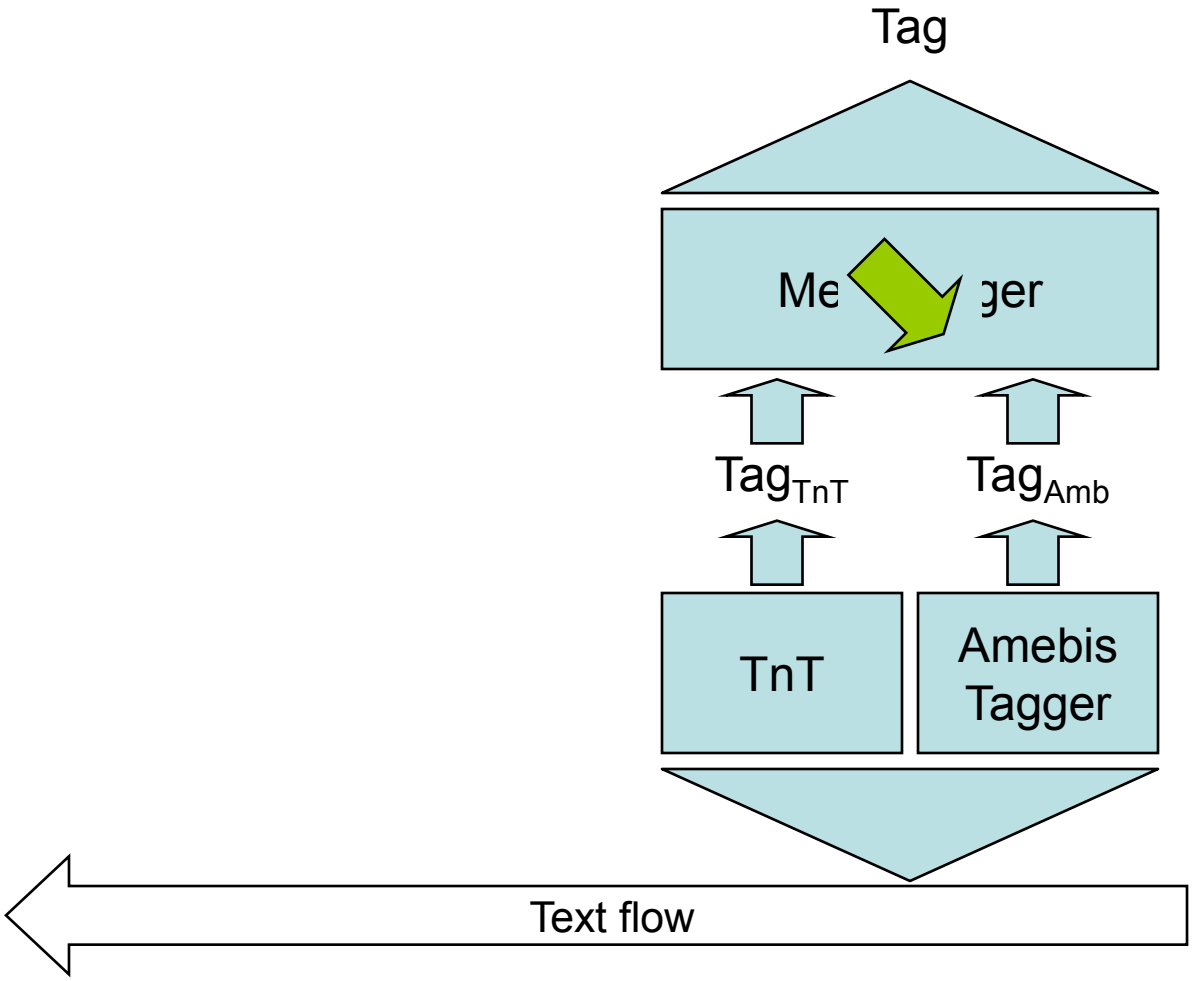
•Purple: TnT correct, Amebis incorrect



Example

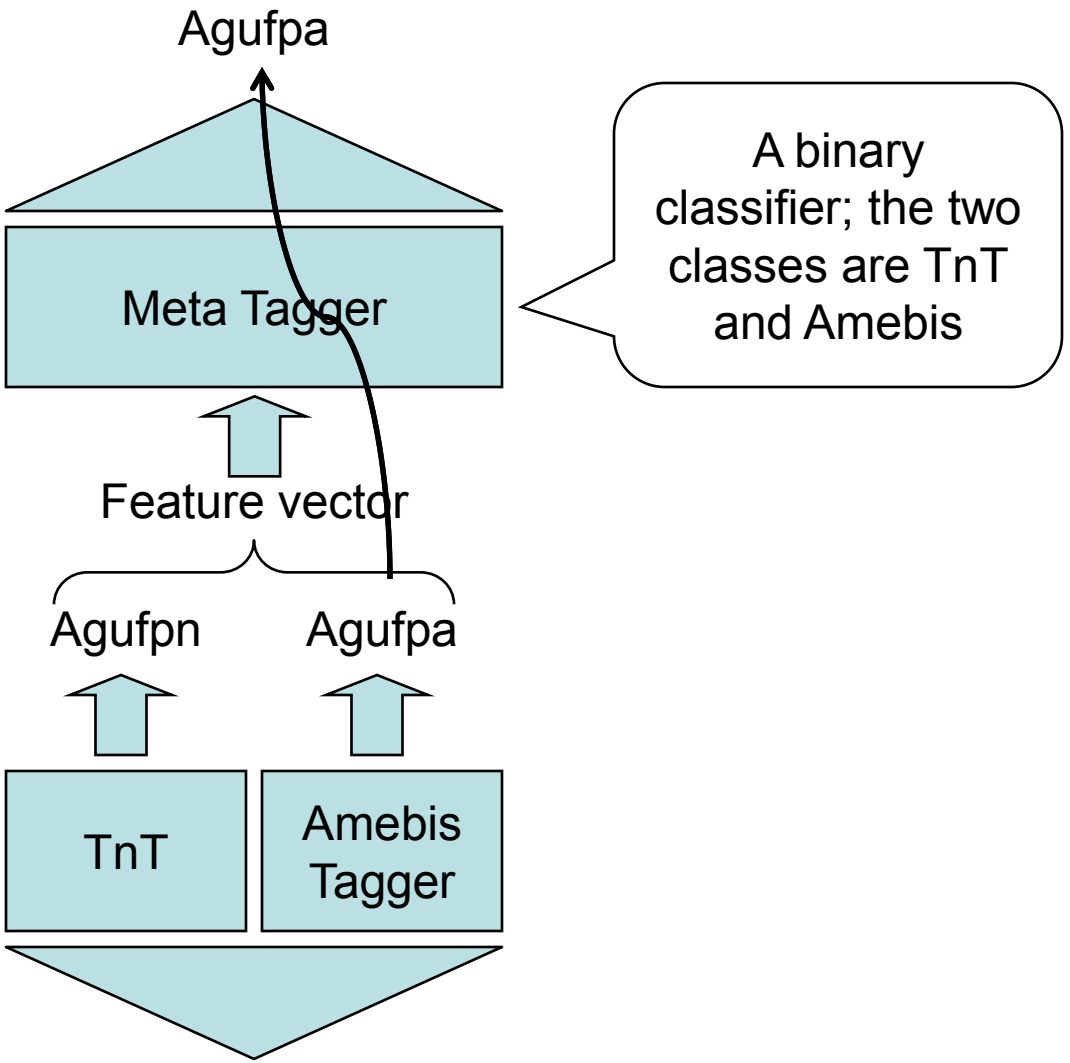
	True	TnT	Amebis
Preiskave	Ncfpn	Ncfpn	Ncfsg
med	Si	Ncmsan	Si
sodnim	Agumsi	Agumpd	Agumsi
postopkom	Ncmsi	Ncmpd	Ncmsi
so	Va-r3p-n	Va-r3p-n	Va-r3p-n
pokazale	Vmep-pf	Vmep-pf	Vmep-pf

Combining the taggers



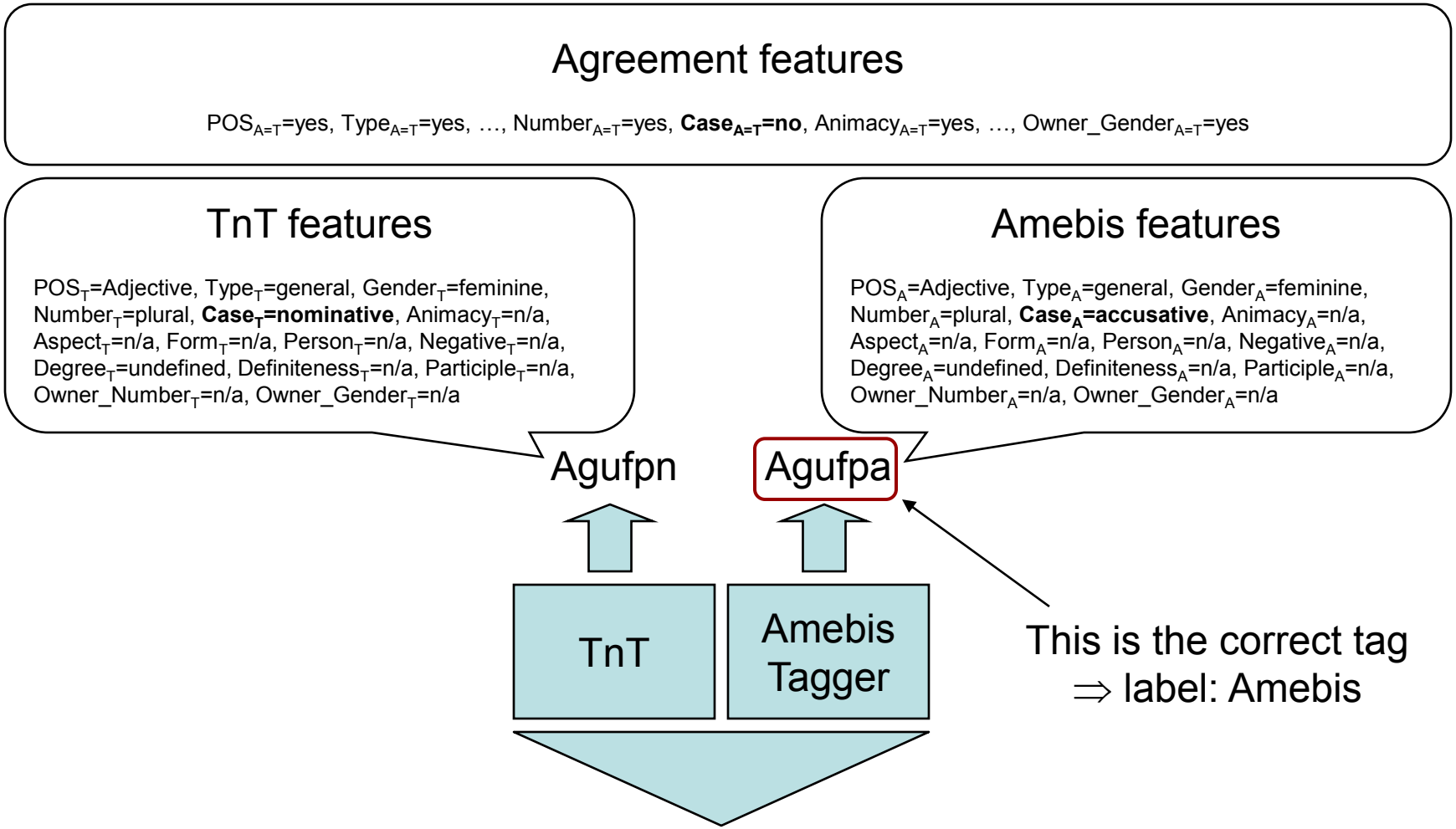
Veža je smrdela po kuhanem ze

Combining the taggers



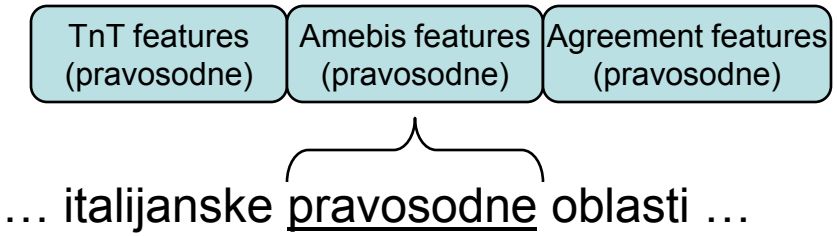
… prepričati italijanske pravosodne oblasti ...

Feature vector construction

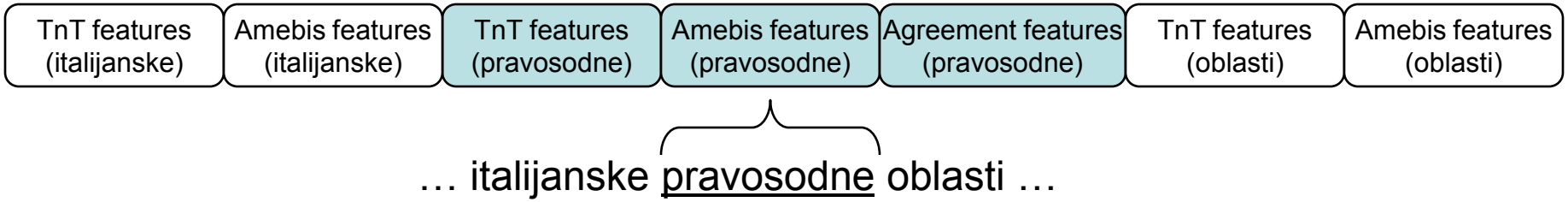


… prepričati italijanske <u>pravosodne</u> oblasti ...

Context



(a) No context



(b) Context

Classifiers

- Naive Bayes
  - Probabilistic classifier
  - Assumes strong independence of features
  - Black-box classifier
- CN2 Rules
  - If-then rule induction
  - Covering algorithm
  - Interpretable model as well as its decisions
- C4.5 Decision Tree
  - Based on information entropy
  - Splitting algorithm
  - Interpretable model as well as its decisions

Experiments: Dataset

- JOS corpus - approximately 250 texts (100k words, 120k if we include punctuation)
- Sampled from a larger corpus FidaPLUS
- TnT trained with 10 fold cross validation, each time training on 9 folds and tagging the remaining fold (for the meta-tagger experiments)

Experiments

- Baseline 1:
  - majority classifier (always predict what TnT predicts)
  - Accuracy: **53%**
- Baseline 2
  - Naive Bayes
  - One feature only: Amebis full MSD
  - Accuracy: **71%**

Baseline 2

- Naive Bayes classifier with one feature (Amebis full MSD) is simplified to counting the occurrences of two events for every MSD *f*:
  - #cases where Amebis predicted the tag *f* and was correct: n_c^f
  - #cases where Amebis predicted the tag *f* and was incorrect n_w^f
  - NB gives us the following rule, given a pair of predictions MSDa and MSDt: if $n_c^{MSDa} < n_w^{MSDa}$ predict MSDt, else predict MSDa.

Experiments: Different classifiers and feature sets

- Classifiers: NB, CN, C4.5
- Feature sets:
  - Full MSD
  - Decomposed MSD, agreement features
  - Basic features subset of the decomposed MSD features set (Category, Type, Number, Gender, Case)
  - Union of all features considered (full + decompositions)
- Scenarios:
  - no context
  - Context, ignore punctuation
  - Context, punctuation

Results

- Context helps
- Punctuation slightly improves classification
- C4.5 with basic features works best

No context

Feature set / Classifier	FULL TAG	DEC	BASIC	FULL+DEC
NB	73.90	67.55	67.50	69.65
C4.5	73.51	**74.70**	74.23	73.59
CN2	60.61	72.57	71.68	70.90

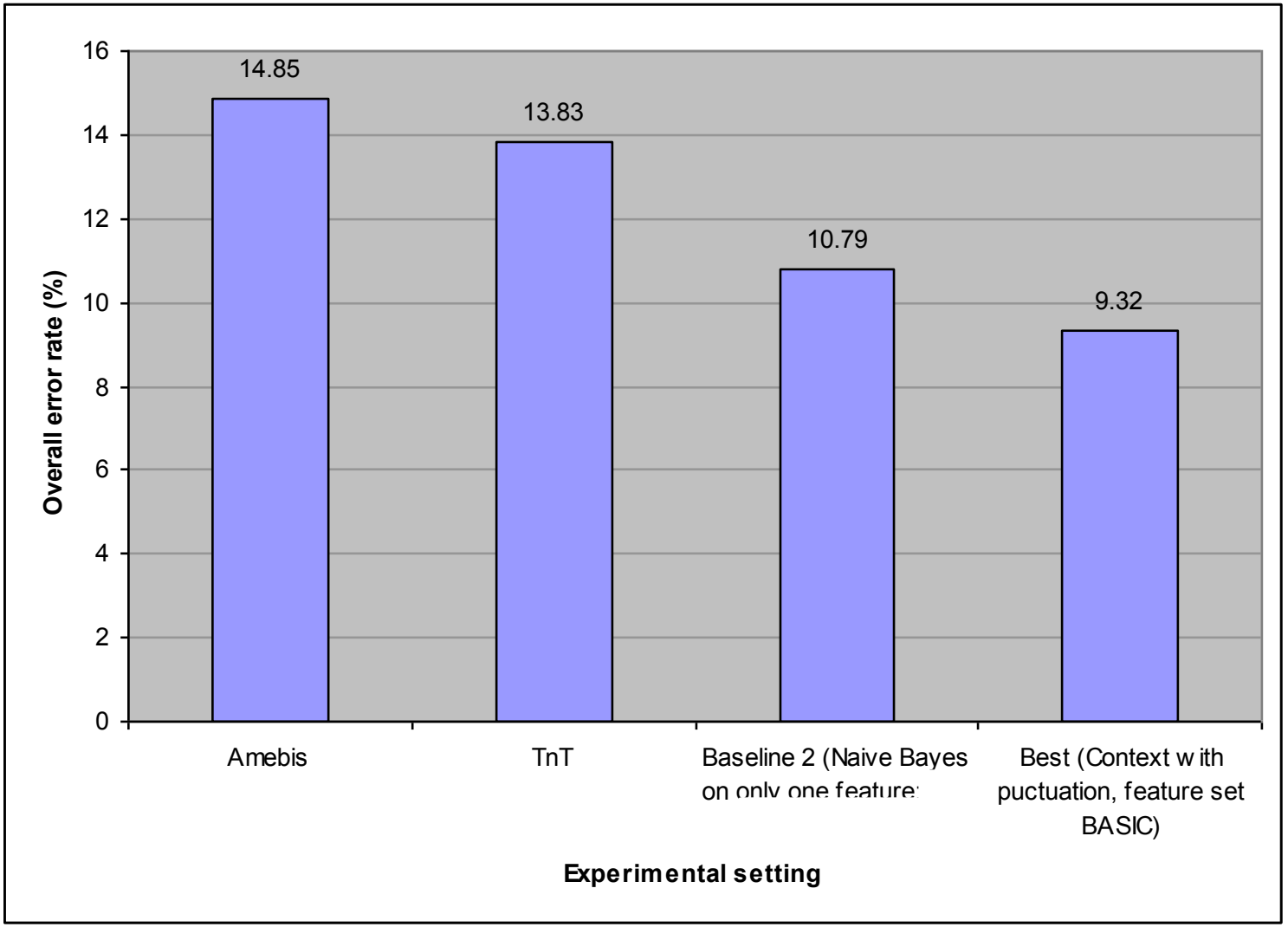
Context without punctuation

Feature set / Classifier	FULL TAG	DEC	BASIC	FULL+DEC
NB	73.10	68.29	67.96	70.55
C4.5	73.10	78.51	**79.23**	76.72
CN2	62.16	73.26	72.75	72.29

Context with punctuation

Feature set / Classifier	FULL TAG	DEC	BASIC	FULL+DEC
NB	73.44	68.32	68.14	70.53
C4.5	74.18	78.91	**79.73**	77.68
CN2	62.23	74.27	72.82	73.01

Overall error rate



Thank you!