

Accurate Stemming and Email Classification

Tanja Gaustad and Gosse Bouma
Alfa-Informatica
University of Groningen
The Netherlands
{tanja|gosse}@let.rug.nl

Main Questions

- Does stemming help in email classification?
- Does more accurate stemming improve results?

Overview

1. Accurate Stemming

- Dutch Porter Stemmer
- Stemmer with Dictionary Lookup
- Application-Independent Evaluation

2. Stemming and Email Classification

- Setup
 - Dataset
 - Classification Algorithm
- Results and Evaluation
 - Quantitative Results
 - Qualitative Evaluation
- Conclusions and Questions

Dutch Porter Stemmer

Porter Stemmer:

- Rule-based suffix stripper
- No lexicon

Porter's Algorithm:

- Series of steps
- Each step removes a certain type of suffix using substitution rules
- Substitution rules only apply when certain conditions hold (e.g. minimal length of resulting stem)

Dutch Porter Stemmer: Kraaij and Pohlman 1994

<http://www-uilots.let.uu.nl/~uplift/>

Examples:

afwachting → <i>afwacht</i>	beste → <i>best</i>
heten → <i>heet</i>	open → <i>oop</i>
uitgelaten → <i>uitlaat</i>	

Stemmer with Dictionary Lookup

Combination of various *existing* resources:

- CELEX for information on stems
- Jan Daciuk's FSA morphology tools for dictionary lookup
- Dutch Porter Stemmer as a backup strategy

Stemmer with Dictionary Lookup (cont.)

Initialisation:

- Extract lists of wordforms and stems from CELEX (381,292 wordforms, 124,136 stems for Dutch)
- Build FSA-dictionary using these lists

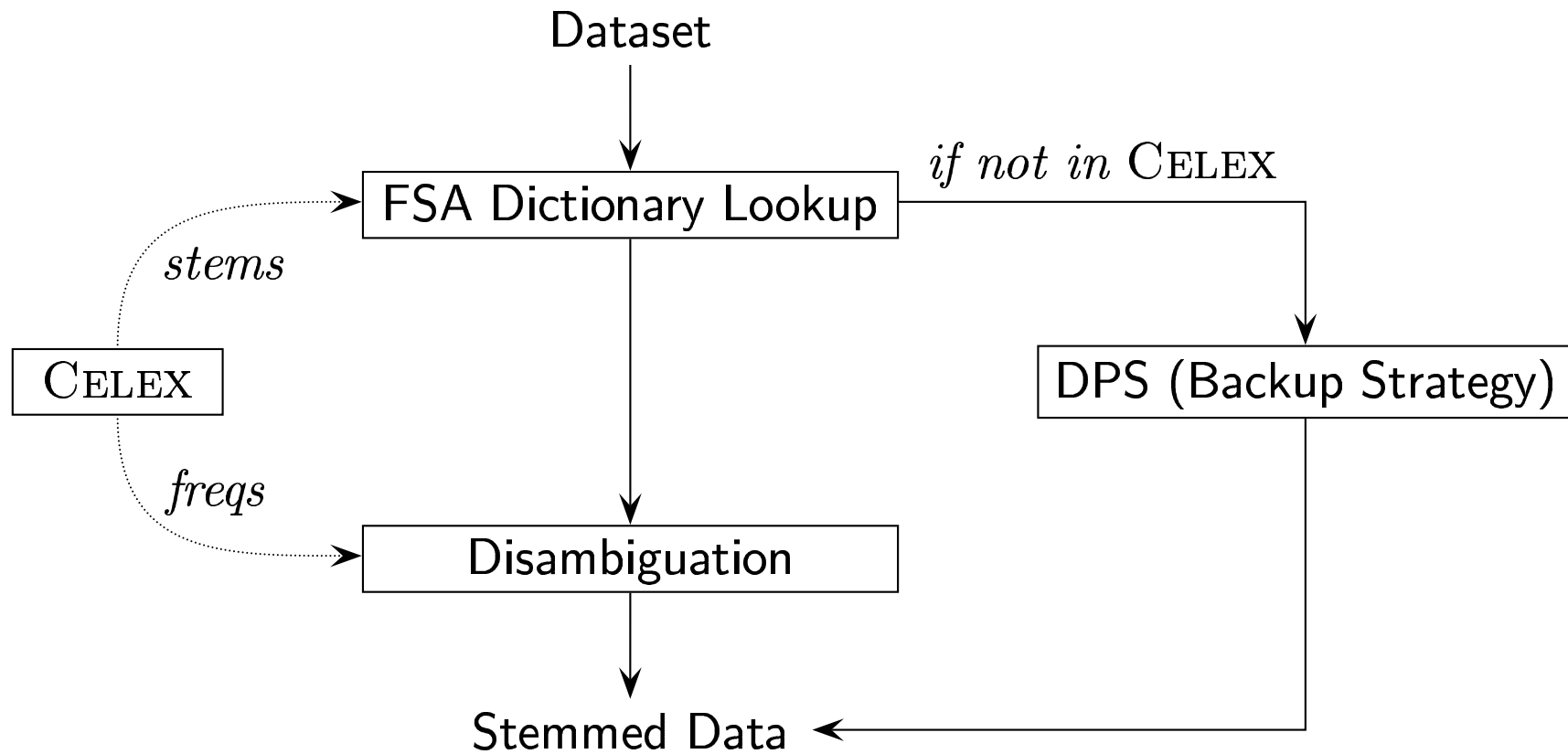
Actual Stemming:

- Perform dictionary lookup in FSA
- Process words not found in CELEX with Dutch Porter Stemmer

Disambiguation:

- Ambiguous output: 58.25% of all wordforms in evaluation file (45,000 words) attributed more than one stem
⇒ Choose most frequent stem (according to CELEX)

Stemmer with Dictionary Lookup (cont.)



Application-Independent Evaluation

Motivation: Assess quality of new stemmer

Corpus: Dutch children's books, ca. 45,000 words,
manually annotated Gold standard

Results:

Stemmer	Accuracy	Time (in s.)
Porter	79.23%	5
SteDL	98.23%	14
SteDL FSA		0.5
SteDL Scripts		13.5
Building FSA		59

- Accuracy: Substantial improvement
- Speed: Not substantially slower

⇒ Dictionary lookup is feasible and helps

Setup

Dataset

Dataset: Email dataset from helpdesk of free internet provider

Total emails	41,000
Total classified emails	5,965
Total categories	293
Non-empty categories	193
Categories >12 emails	69
Total emails in 69 categories	5,519
Average length	79 words

Setup

Classification Algorithm

Software: BOW Software package (IR and document classification)

Settings:

- Statistical classification model: Naive Bayes
- Unigram and bigram word counts
- Stoplist (+ preprocessing, e.g. tokenizing)
- 90% training data, 10% test data
- 10-fold crossvalidation

Feature space reduction:

- Pruning by word occurrence: > 15
- Information gain: 2,500 most informative words

Results and Evaluation

Quantitative Results

N-best classification: Not only first best (n=1) result, but n-best results
⇒ increases chance of correct classification
⇒ acceptable in call center environment (human agent)

Results:

Emails	n	Unstemmed av. (stderr)	Porter av. (stderr)	SteDL av. (stderr)	Base.
>12	1	41.87 (0.43)	41.43 (0.36)	40.53 (0.47)	15.93
	3	68.97 (0.45)	67.88 (0.46)	68.06 (0.63)	26.87
	5	78.35 (0.26)	78.24 (0.37)	78.60 (0.34)	35.04

Results and Evaluation

Qualitative Evaluation

Expectation: Improvement of results through generalization (words occur less sparsely)

Actual Generalization: In top 100 uni- and bigrams only **4** stems combination of different word forms
⇒ not as much generalization as expected

Number of features: Feature set not substantially smaller with stemming

	unigrams	bigrams
Unstemmed	24,568	169,870
Porter	23,709	163,104
SteDL	23,347	156,407

- SteDL reduces features by 5%-8%.

Conclusions and Questions

Conclusion: Combination of accurate techniques does not automatically yield better results

Questions:

- Specific for **email classification**?
⇒ No improvement on Dutch newspaper corpus with stemming (own experiments and Spitters, 2000)
- Specific for **Dutch**?
⇒ Stemming improves performance for German email classification (Busemann et al., 2000)

Conclusions and Questions (cont.)

Email classification specific?

Experiments on newspaper corpus (Volkskrant):

	unigrams	bigrams
Unstemmed	61,721	538,439
Porter	52,313	504,350
SteDL	50,850	478,928

- SteDL reduces number of features with 10-20%.

Settings	Unstemmed av. (stderr)	Porter av. (stderr)	SteDL av. (stderr)
uni/IG=10K	87.42 (0.18)	86.73 (0.19)	87.28 (0.22)
bi/IG=10K	87.40 (0.21)	87.19 (0.18)	87.64 (0.20)
bi/IG=10K occurs=5	87.55 (0.21)	86.99 (0.21)	87.31 (0.19)

- SteDL performs better than Porter, but not significantly better than Unstemmed.

⇒ No improvement on a different corpus.