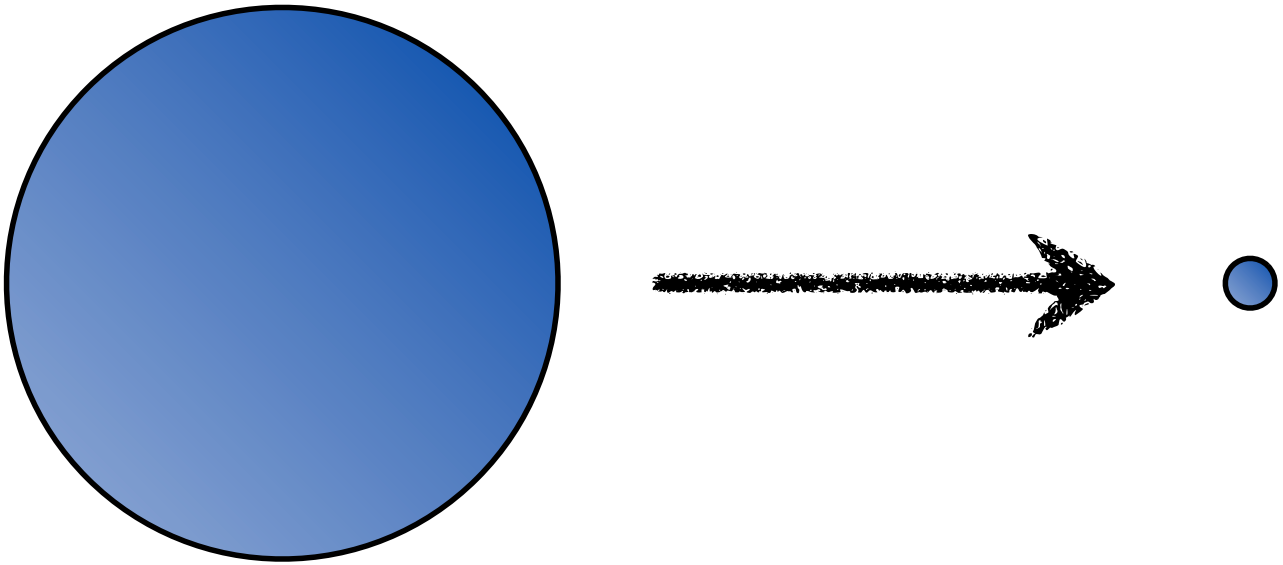
Feature selection for fluency ranking

Daniël de Kok

INLG 2010, July 7-9 2010



Model

Compressed model

Overview

- Motivation
- Feature selection methods
- Experimental setup
- Results
- Conclusions

Characterizing fluency

How to characterize fluency of a sentence?

- *wij gaan dieper op die vraag in*
- *wij gaan dieper in op die vraag*
- *dieper gaan wij op die vraag in*
- *op die vraag gaan dieper wij in*

(Translated: *we will discuss that question in more depth*)

Characterizing fluency

- N-gram models
- Feature-based models, such as maximum entropy models and support vector machines outperform output-based models:
  - Nakanishi et al., 2005
  - Velldal and Oepen, 2006
  - Velldal, 2008
- What features should we use?
  - Feature engineering
  - Very generic templates
- Can we gain insights via the second approach?

Generic templates

- Good performance can be achieved by using very generic feature templates (Velldal and Oepen, 2006)
- Leads to opaque models:
  - Large number of features
  - (Nearly) identical features
- Can we find small and transparent models with relatively little labor?

Feature selection

- Feature selection tries to extract $S \subset F$ from a set of features F
- Model based on S should perform comparable to a model based on F
- Particularly useful iff $|S| \ll |F|$.
- Previous work:
  - Frequency cut-offs (Ratnaparkhi, 1999)
  - Maximum entropy selection for classification (Berger et al. 1996)
  - Selection ℓ_1 regularization (Perkins, et al. 2003)

Why would feature selection work?

- Some features only change sporadically in value for different realizations of an input
- Some features correlate strongly with other features (show comparable behavior)
- Some features have little or no correlation with the classification or ranking.

Frequency-based selection

- Count how often a feature value changes within a given context (within the realizations of an input)
- Order features by this count
- Variation: exclude features that change of value in less than *n* contexts
- Can not detect feature overlap, or noisy features.

Correlation-based selection

- Start with the ordering imposed by frequency-based selection.
- Consider features one by one, selecting features that do not show a high correlation with a previously selected feature (sample correlation coefficient)
- Cannot detect noisy features - no correlation with selected features

Maximum entropy selection

- Start with a uniform model (assigning the same probability to each realization)
- Add the feature to the model that gives the highest improvement of prediction of the training data
- Obey the principle of maximum entropy
- Assume that the weights of features already in the model do not change by the addition of a feature: only optimization of the weight of the candidate feature required

Maximum entropy selection (2)

- Maximum entropy selection as described by Berger et al, 1996 and Zhou et al. 2003
- Modified for ranking tasks, rather than classification tasks
- Mathematical details: see paper

Task

- Fluency ranker for a Dutch sentence realizer based on the Alpino system (Van Noord, 2006)
- Generation from dependency structures
- Select the most probable realization given the dependency structure

Output features

Auxiliary distributions:

- Word trigram model
- Tag trigram model

Construction features

- Features from parse disambiguation:
  - Topicalization of (non-)NP and subjects
  - Use of long-distance/local dependencies
  - Orderings in the middle field
  - Identifiers of grammar rules used to build the derivation tree
  - Parent-daughter combinations
- Features described by Velldal (2008):
  - Local derivation subtrees with optional grand-parenting
  - Local derivation subtrees with back-off and optional grand-parenting
  - Binned word domination frequencies of the daughters of a node

Evaluation/training data

- Dependency structures constructed by parsing 11764 random (unannotated) Wikipedia sentences of 5-25 tokens
- Best parse considered the correct parse (approx. 90% concept accuracy)
- Original sentence considered the best realization
- Realizations and their derivation trees generated using the Alpino chart generator
- Training and testing data obtained by extracting features from each realization
- Quality of the realization is estimated by comparing the realization with the original sentence using the General Text Matched method (Melamed, et al. 2003)

Methodology

Training (5884 dependency structures):

1. Randomly select 100 training instances for every dependency structure in the training data.
2. Apply feature selection methods, selecting 100..5000 features with steps of 100.
3. Train a maximum entropy model for each set of features

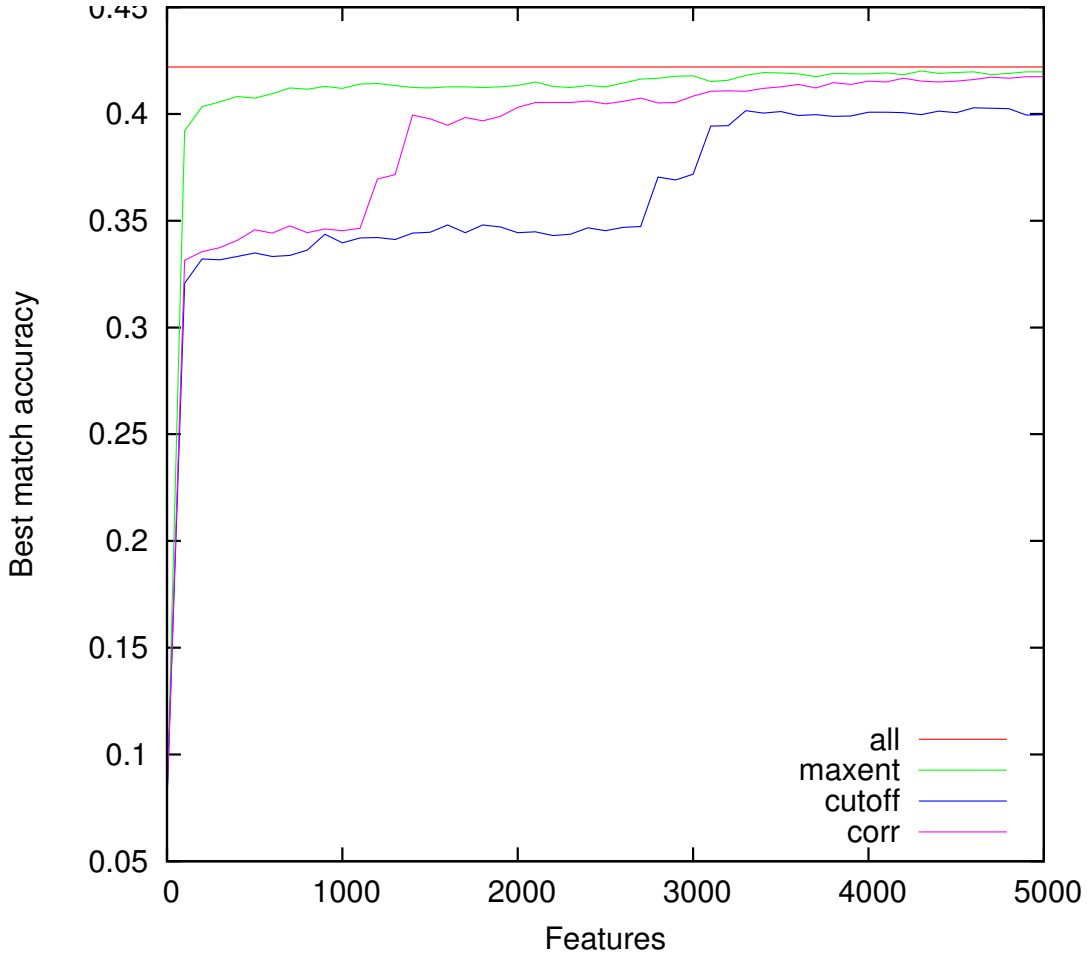
Evaluation (5880 dependency structures)

- Select dependency structures with 5 or more realizations
- For every dependency structure, select the realization that is the closest to the correct realization (according to the General Text Matcher method).
- Calculate the fraction of instances for which the model picked the correct realization (best match accuracy)

Results (best)

Method	Features	Accuracy
Random	0	0.0778
Tag n-gram	1	0.2039
Word n-gram	1	0.2799
Word/tag n-gram	2	0.2908
All	490667	0.4220
Fixed cutoff (4)	90103	0.4181
Frequency	4600	0.4029
Correlation	4700	0.4172
Maxent	4300	0.4201

Selection methods



Most effective features

1. Word trigram model
2. Tag trigram model
3. Placement of the predicative complement after the copula:
   - *Amsterdam is de hoofdstad van Nederland* (*Amsterdam is the capital of The Netherlands*)
   - *De hoofdstad van Nederland is Amsterdam* (*The capital of The Netherlands is Amsterdam*)
4. Dispreference of topicalized non-subject NPs:
   - *Jan eet de soep* (*Jan eats the soup*)
   - *de soep eet Jan* (*the soup eats Jan*)

Most effective features (2)

5. Prepositional complements that are not topicalized:
   - *dit zorgde voor veel verdeeldheid* (*this caused lots of discord*)
   - *voor veel verdeeldheid zorgde dit* (*lots of discord caused this*)
6. PP-modifiers following the head in conjuncts:
   - *groepen van bestaan of khandas* (*planes of existance or khandas*
   - *van bestaan groepen of khandas* (*of existance planes or khandas*)

Most effective features (3)

7. Topicalized PP if the PP modifies a copula in a subject-predicate structure:
   - *volgens Williamson is dit de synthese* (*according to Williamson is this the synthesis*)
   - *dit is de synthese volgens Williamson* (*this is the synthesis according to Williamson*)

8-10. Preferences involving punctuation:
   - *Bill Clinton - een man zonder angsten* (*Bill Clinton - a man without fears*)
   - *een man zonder angsten - Bill Clinton* (*een man zonder angsten Bill Clinton*)

Conclusions

- Fluency models can be compressed enormously by applying feature selection
- The maximum entropy feature selection method shows a high accuracy after selecting just a few features
- The commonly used frequency-based selection method requires the selection of far more features to achieve a comparable performance
- Correlation-based selection shows that the ineffectiveness of frequency-based selection can be explained partly by overlap
- Hopefully, feature selection will give us insights for developing more targeted features.

Thank you!

Software implementing these feature selection methods is available from: http://danieldk.eu/Code/FeatureSqueeze/

Thank you!