Discriminative Features in Stochastic Attribute-Value Grammars

Daniel de Kok
Computational Linguistics, Center for Language and Cognition, University of Groningen, The Netherlands

Research goals

Reversible Stochastic Attribute-Value Grammars (de Kok et al., 2011) use one maximum entropy model for parse disambiguation and fluency ranking.

RSAVGs rely on the premise that preferences are shared between parsing and generation. For instance, in Dutch, preferences with respect to subject fronting ought to be shared between production and comprehension to make communication effective.

In this work we show that RSAVGs do indeed use task-independent features.

Methodology

1. Find an effective feature selection method for RSAVGs.
2. Use this method to select the most effective features in reversible and directional models.
3. For each model, find per-class feature contributions.

Candidates

We compared three iterative feature selection models:

- **Gain-informed selection**: select the feature with the highest contribution to the model and perform a parameter optimization.
- **Grafting**: select the feature with the highest gradient given the model and perform a parameter optimization.
- **Grafting-light**: select the feature with the highest gradient given the model and perform one iteration of parameter optimization.

Experimental setup and evaluation

- Experiments performed using the Alpino parser and generator for Dutch (van Noord, 2006; de Kok and van Noord, 2010).
- Training data: Alpino Treebank (cdbl part) (van der Beek et al., 2002).
- Evaluation data: Part of Trouw 2001 from LASSY (van Noord et al., 2010).
- The following classes of features are used:
  - **Parse disambiguation:** lexical frame selection and dependency triples.
  - **Fluency ranking:** trigram language models.
  - **Both directions:** rule identifiers and other syntactic features.
- Let each feature selection method select 1800 features, evaluate models with 0..1800 features.

Comparing the candidates

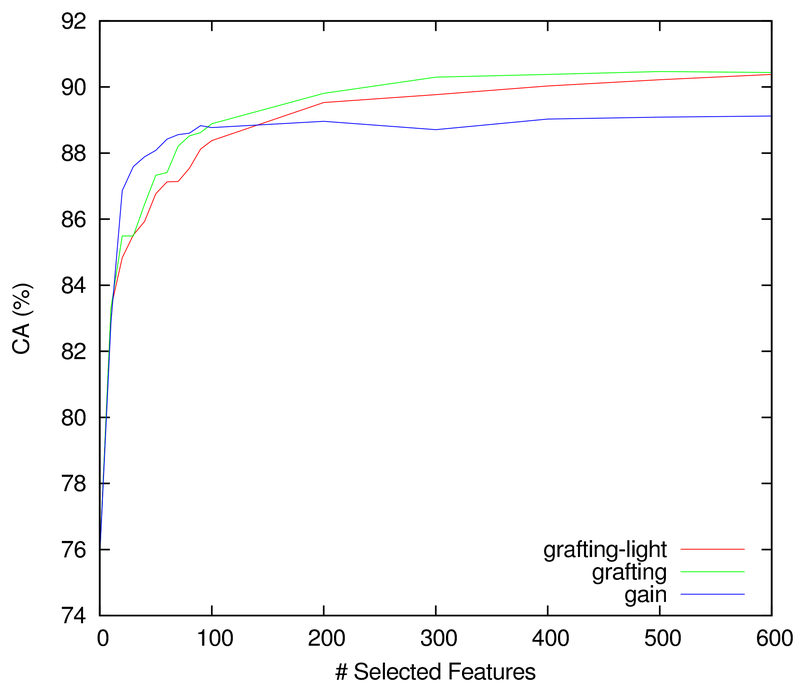

Figure 1: *Application of feature selection methods to parse disambiguation.*

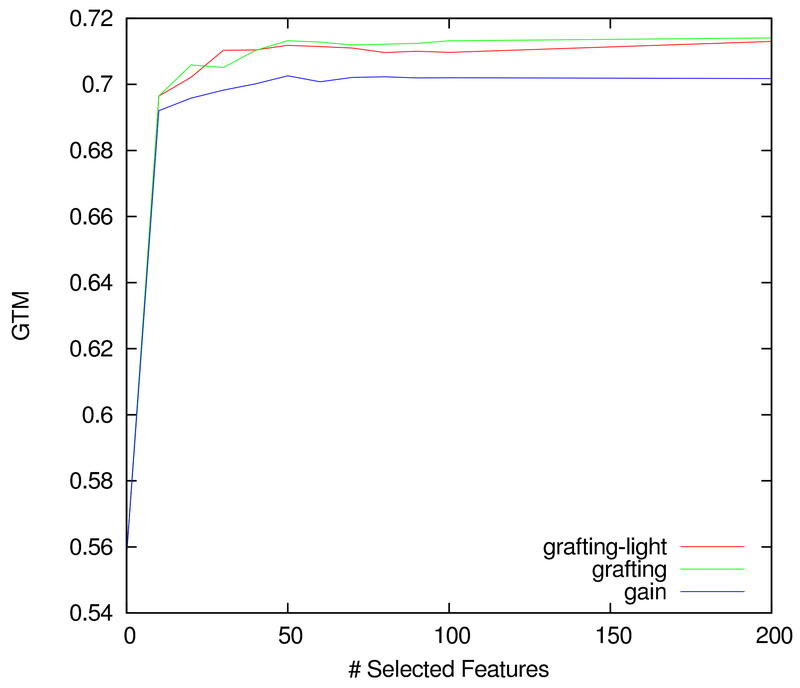

Figure 2: *Effectiveness of feature selection methods in fluency ranking.*

For both tasks, grafting is the most effective selection method.

Contribution of features in RSAVGs

We calculate the contribution of a feature f_i according to the evaluation function e, where $F_{0..i}$ is a model trained with the i most discriminative features, F_0 the uniform model, and $n = 300$.

$$c(f_i) = \frac{e(F_{0..i}) - e(F_{0..i-1})}{e(F_{0..n}) - e(F_0)}$$

Class	Directional	Reversible
Dependency	21.53	13.35
Lexical	33.68	32.62
N-gram	0.00	0.00
Rule	37.61	47.35
Syntactic	7.04	6.26

Table 1: *Per-class contribution to the improvement of the model over the base baseline in parse disambiguation.*

Class	Directional	Reversible
Dependency	0.00	0.00
Lexical	0.00	0.00
N-gram	81.39	79.89
Rule	14.15	15.75
Syntactic	3.66	4.39

Table 2: *Per-class contribution to the improvement of the model over the baseline in fluency ranking.*

Conclusions

- Grafting is the most effective feature selection method of the candidates. Grafting-light provides comparable performance, but is multiple times faster.
- Models can be compressed massively using feature selection with very little loss of accuracy.
- RSAVGs do rely on features used in both directions, even more than directional models..

Software

- Alpino dependency parser and generator:
  `http://www.let.rug.nl/vannoord/alp/Alpino/`
- TinyEst maximum entropy parameter estimator:
  `http://github.com/danieldk/tinyest`