Feature Selection for Cross-Linguistic Parse Ranking

Comparison of Parse Ranking Accuracy for English and Japanese

W.P. McNeill

DELPH-IN Barcelona Summit

July 21, 2009

Table of Contents

Research Questions

- What are the parse selection accuracies in different languages for different feature sets?
- Do the same feature sets work for different languages?

Languages and Features

Languages

- English — jhpstg corpus, ERG grammar
- Japanese — Tanaka corpus, JACY grammar

Features

- Grandparenting — 0
- Active Edges — true, false
- Constituent Weight — 1, 2, 0
- N-gram — 3,4
- N-gram Backoff — true, false

Evaluation Metric

1-best exact match

Raw Results

Fixed dimensions

- Grandparenting 0
- Relative Tolerance 1×10^{-8}
- Variance 1

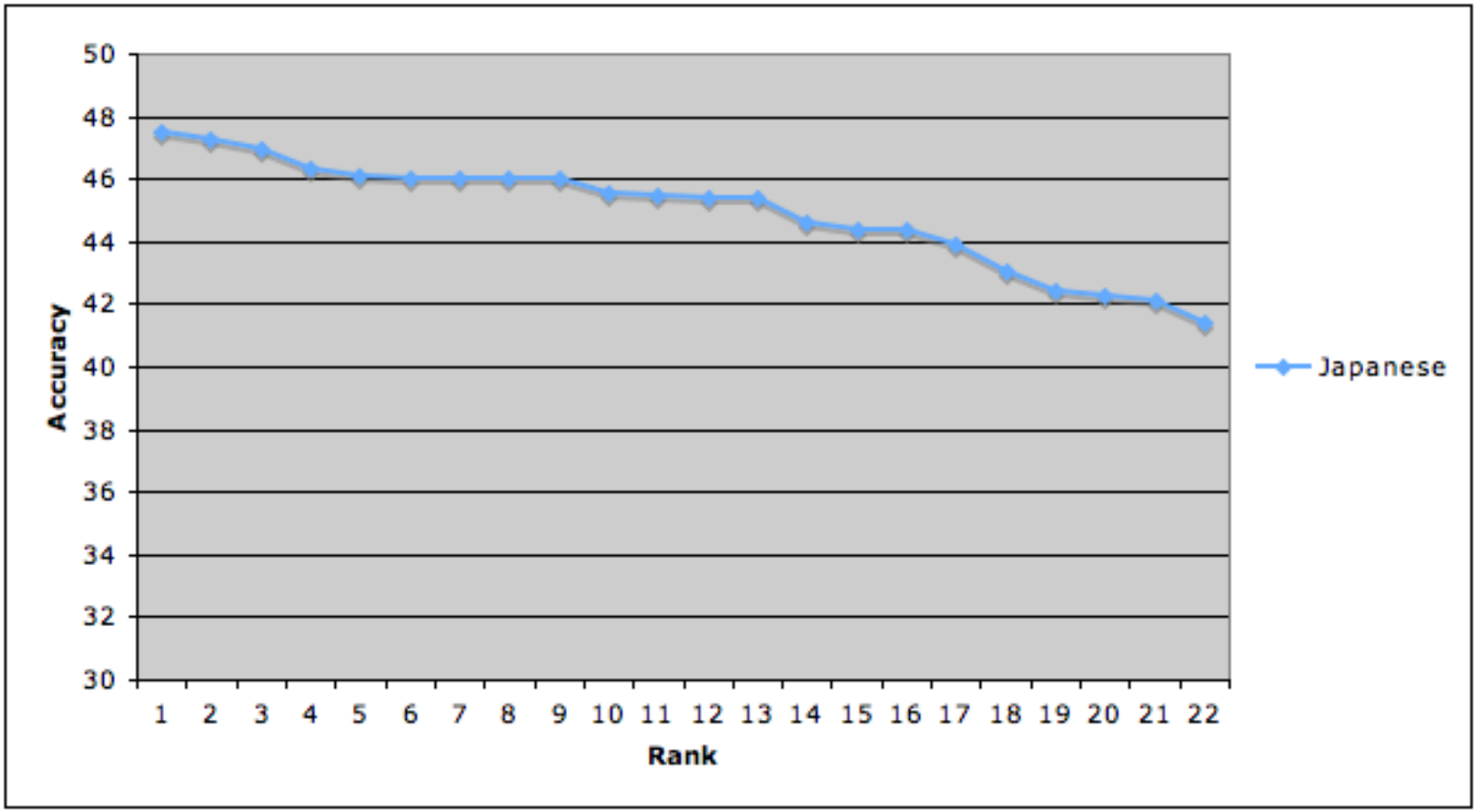
Best Results

- Japanese **47.46**
- CW=2, AE=false, **N-gram=3**, N-gram Backoff = true
- English **37.13**
- CW=2, AE=false, **N-gram=4**, N-gram Backoff = true

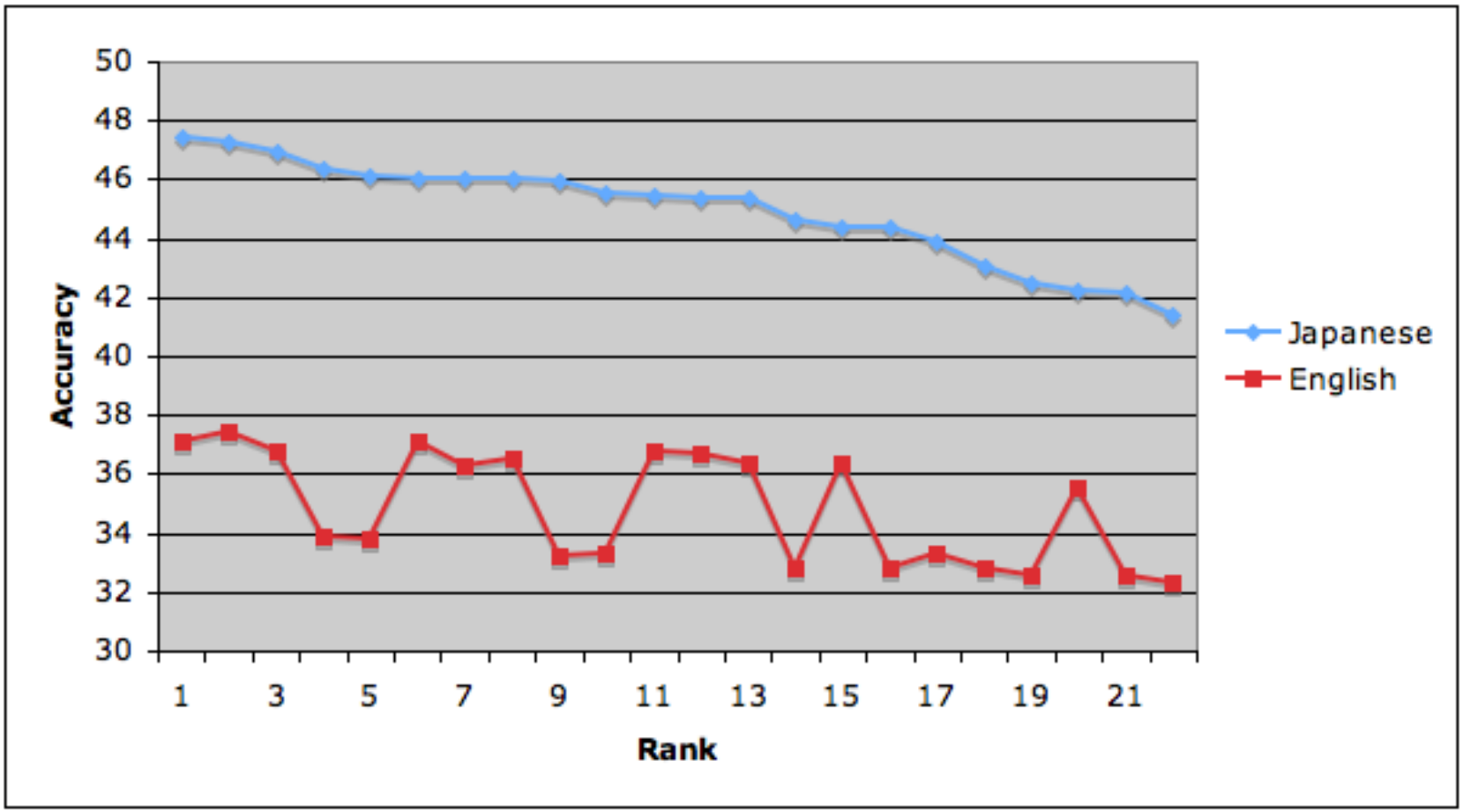
Cross-Linguistic Feature Sets

- For a single language just pick the best feature set
- How do you quantitatively navigate the feature space for more than one language?
- Which feature subsets make two languages the most dissimilar?

Japanese and English Accuracy



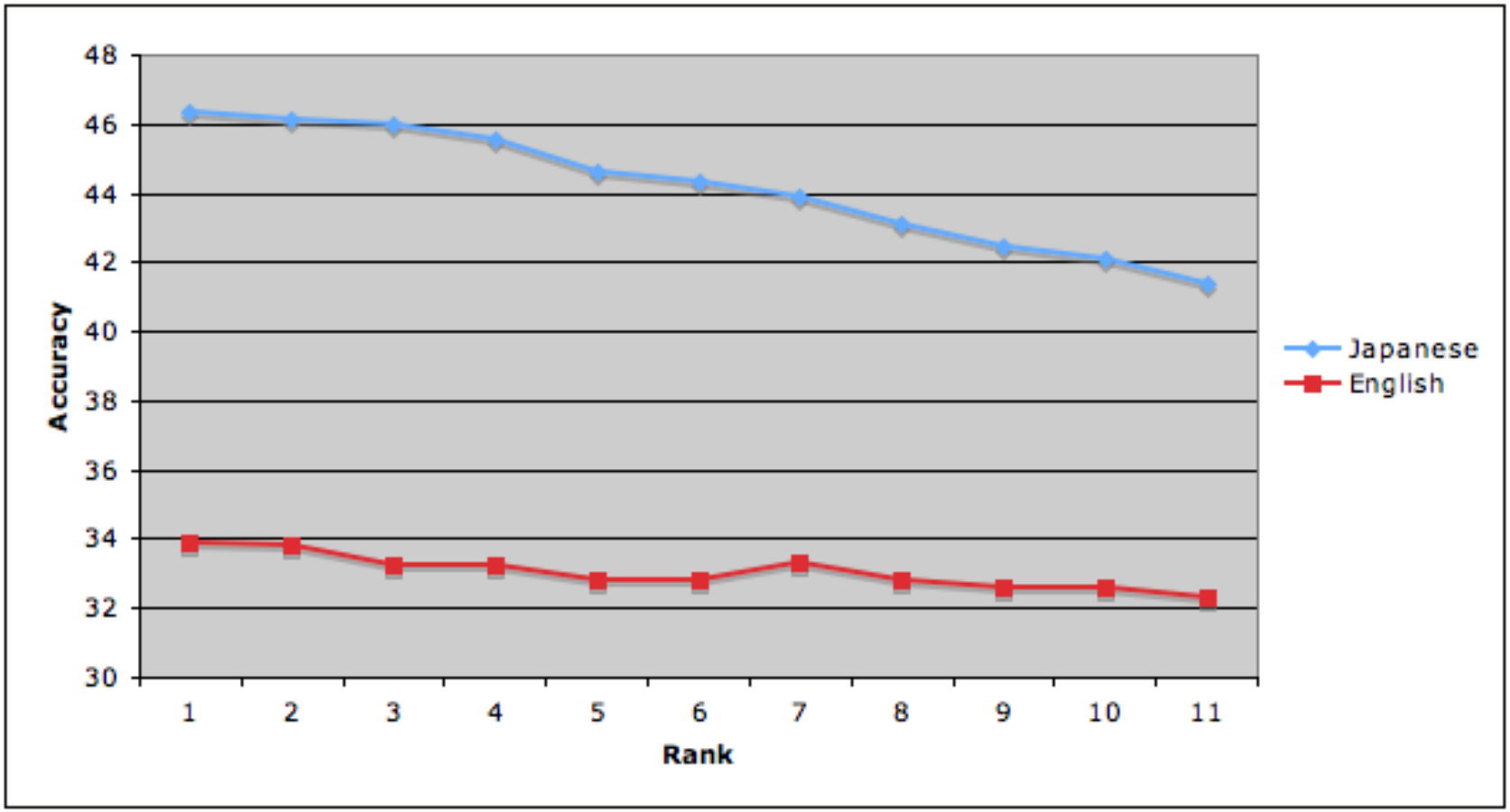
Japanese and English Accuracy



Relative Monotonicity

- Japanese is monotonically decreasing by construction
- Which data points can I remove from English to make it also monotonic?

Japanese and English Accuracy, Active Edges = false



Quantifying Correlation

- Use Pearson's Rank Correlation Coefficient

$$\rho = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}}$$

  where x_i and y_i are corresponding rankings
- ρ ranges from -1 (anticorrelated) to $+1$ (correlated)
- $\rho = 0$ is uncorrelated

Discussion

Feature Exclude	Pearson's Correlation	p-value
Active Edges=false	0.89	0.0003
Active Edges=true	0.87	0.0005
N-gram backoff = true	0.85	0.0010
N-gram backoff = false	0.80	0.0032
N-gram=0	0.71	0.0003
CW=0	0.69	0.0098
None	0.63	0.0015
CW=1	0.63	0.0121
CW=2	0.62	0.0097
N-gram=4	0.62	0.0421
N-gram=3	0.53	0.0746

Preliminary Conclusions

- Mostly the same feature set performs equally well for Japanese and English
- Methodology for extracting most discriminative features uses correlation coefficient
- Most discriminative English/Japanese feature is Active Edges

Future Work

- Finish generating grid points
- Test stability on different iterations and fold numbers
- Different accuracy metrics