Authorship Attribution: Using Rich Linguistic Features when Training Data is Scarce

Ludovic Tanguy, Franck Sajous, Basilio Calderone and Nabil Hathout

CLLE-ERSS: CNRS & University of Toulouse, France

PAN 2012 – Authorship Attribution - CLEF

- General method for all subtasks
  - Maximum Entropy classifier (csvLearner)
  - Substantial effort in feature engineering
    - *Many linguistically rich features*
  - No feature selection
  - Whole texts as items (no splitting)
- Four runs were submitted:
  - Run 1 (CLLE-ERSS1): char. trigrams + all linguistic features
  - Run 2 (CLLE-ERSS2): character trigrams only
  - Run 3 (CLLE-ERSS3): bag of words (lemma frequencies)
  - Run 4 (CLLE-ERSS4): a selection of 60 synthetic features

- All training and test texts were :
  - Normalised for encoding
  - De-hyphenised (based on a lexicon)
  - POS-tagged and parsed (Stanford CoreNLP)

- No split?
  - Using splits of the same few texts is misleading (textual cohesion)
  - No cross-validation data available…

List of features (1)

- Contracted forms
  - Average ratio of frequencies (« *do not* » vs « *don't* », etc.)
- Phrasal verbs
  - Frequency of all verb-prepositions pairs (« *put on* », etc.)
- Lexical genericity and ambiguity
  - Average depth in WordNet
  - Average number of synsets per word
- Frequency of POS trigrams
- Syntactic dependencies
  - Frequency of all word-relation-word triples (« *cat* – *subj* – *eat* »
- Syntactic complexity
  - Average depth of syntactic parse trees
  - Average length of syntactic links

List of features (2)

- Lexical cohesion
  - Density of semantically-similar word pairs
    - *(according to Distributional Memory database)*
- Morphological complexity
  - Frequency of suffixed words
- Lexical absolute frequency
  - Repartition of words according to Nation’s wordlists
- Punctuation and case
  - Frequency of punctuation marks
  - Frequency of uppercased words
- Direct speech
  - Ratio of sentences between quotes
- First person narrative
  - Relative frequency of « I » (per verb, outside quotes)

- Closed-class tasks (A,C,I)
  - Choose the author with highest probability
- Open-class tasks (B,D,J)
  - Author is « unknown » if

$$\max(p) < \text{mean}(p) + 1.25 * \text{st.dev}(p)$$

- Results :
  - Overall:
    - *All rich+3char > synthetic rich > lemmas > 3char*
  - Results :
    - *Good for A, I and J*
    - *Average for B*
    - *Bad for C and D*

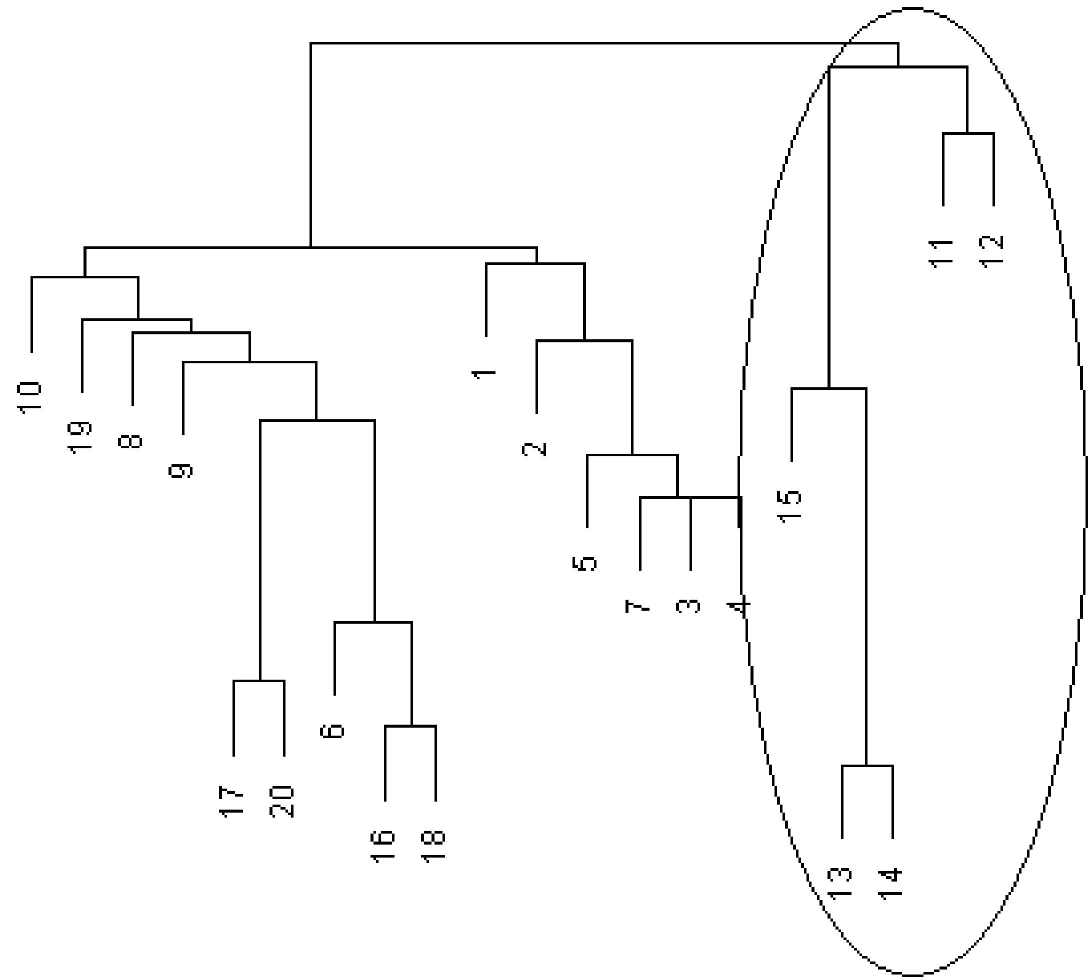
- Lesion studies on test data for tasks A and C
  - Measuring accuracy with different combinations of features
  - Average accuracy gain when adding each subset

Feature Subset	Gain for task A	Gain for task C
Punctuation & case	**+0.204**	**-0.040**
Suffix frequency	+0.097	+0.009
Absolute lexical frequency	**+0.030**	**-0.003**
Syntactic complexity	+0.015	+0.006
Ambiguity/genericity	+0.012	+0.008
Lexical cohesion	+0.002	-0.000
Phrasal verbs (synthetic)	-0.000	+0.022
Morphological complexity	-0.005	-0.002
Phrasal verbs (detail)	-0.006	-0.006
Contractions	**-0.014**	**+0.018**
First/third person narrative	-0.027	-0.026
POS trigrams	**-0.028**	**+0.045**
Char. trigrams	**-0.034**	**+0.206**
Syntactic dependencies	**-0.059**	**+0.089**

$r = -0.48$

- Using MaxEnt as an *unsupervised* classifier
  - Method proposed by DePauw and Wagacha, 2008
- Principles:
  - Training: all paragraphs as training items
    - *Class value = paragraph ID*
  - Reclassifying: every paragraph processed by the trained classifier
    - *Result = square matrix of probabilities (Mp)*
    - *Distance matrix between paragraphs: Md= -log(Mp)*
  - Clustering: regroup similar paragraphs
    - *Hierarchical ascending clustering on Md*
  - Result: highest level clusters

- Task F, Text 4, Run CLLE-ERSS1 (correct guess)



- Conclusions
  - Average results for traditional tasks, quite disappointing
  - Good results for paragraph intrusions
  - Overall, rich features are once more proven to be an improvement over character trigrams
  - There’s still room for improvement with feature selection
    - *Feature efficiency varies greatly across tasks and authors*
    - *Very small linguistic feature subsets can be sufficient*