Analysis of Linguistic Variation

Jefrey Lijffijt

Aalto University, Finland

Abstract

- Many medium to large text corpora have been compiled and annotated
- This enables the study of more diverse and detailed aspects of language
  - E.g., differences between writing style of various age groups/gender/media
- New computational and statistical challenges arise
  - See the various parts of this poster

Profile

- Doctoral student in the group of Heikki Mannila
- Member of ALGODAN, HIIT, PASCAL2
- Research interests include analysis of sequential data and mining bursty patterns
- E-mail: jefrey.lijffijt@aalto.fi

Burstiness

- In linguistics it is often assumed that all words in a corpus are independent
- It has been argued that this is not problematic when there are many samples
- Figure 1 shows how false this statement is
- This effect is known as *burstiness* [2]

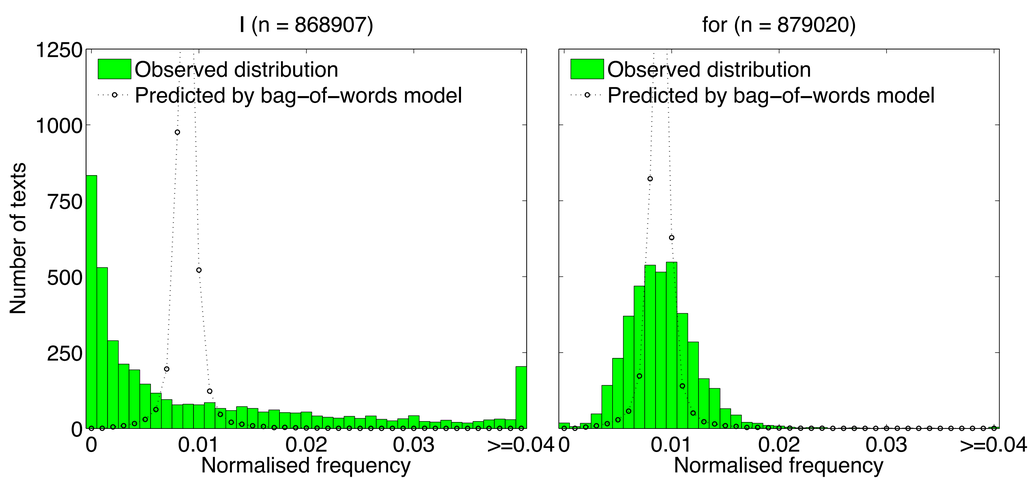

Figure 1: Frequency histograms of the words *I* and *for* in the British National Corpus. The distributions predicted under the bag-of-words assumption are very poor. The pronoun *I* is much burstier than the grammatical word *for* ; the Weibull shape parameter β is 0.57 and 0.93, for *I* and *for* respectively, see the paragraph below. Adapted figure from [3].

Classification of text genres

- Models for genre classification are complex and difficult to interpret
- It appears the main genres (of British English) can be recognized using a simple model and easy to compute surface level features

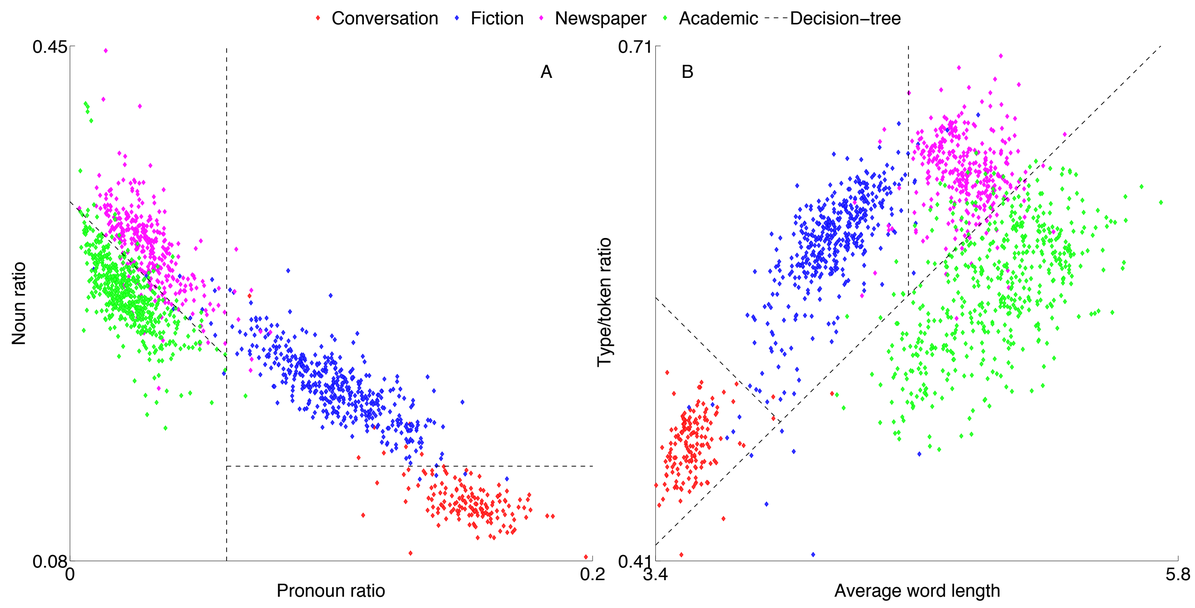

Figure 3: Two models for classification of the main genres of British English. The model was trained using the C4.5 algorithm on the British National Corpus, using both the original features and their cross-terms.

Inter-arrival times

- An *inter-arrival time* of a word is the number of words between two consecutive occurrences

" Finnair believes that it will be able to resume its scheduled service to **and** from New York on Monday, after two days of cancellations caused by hurricane Irene. All three airports serving New York City have been closed because of the hurricane **and** Finnair was forced to cancel flights on Saturday **and** Sunday. The airline is not certain when its scheduled service can be resumed, but the assumption is that Monday afternoon's flight from Helsinki will depart. Some Finnair passengers whose final destination is not New York have been rerouted **and** some have delayed travel plans. The company has also offered ticket holders a refund. *YLE* "

- $IAT_{and} = \{29, 9, 39, 29\}$
- The distribution of inter-arrival times describes the burstiness of a word
- A summary is obtained by fitting a Weibull distribution [1]

Comparison of word frequencies

- We can use statistical testing to find significant variations in writing styles
  - I.e., between time periods, between people or between text types
- Tests commonly employed are based on the bag-of-words assumption (χ^2-test)
- *Burstiness* leads to over-estimation of the significance [3]
- Improved tests based on inter-arrival times or bootstrapping are proposed [3]

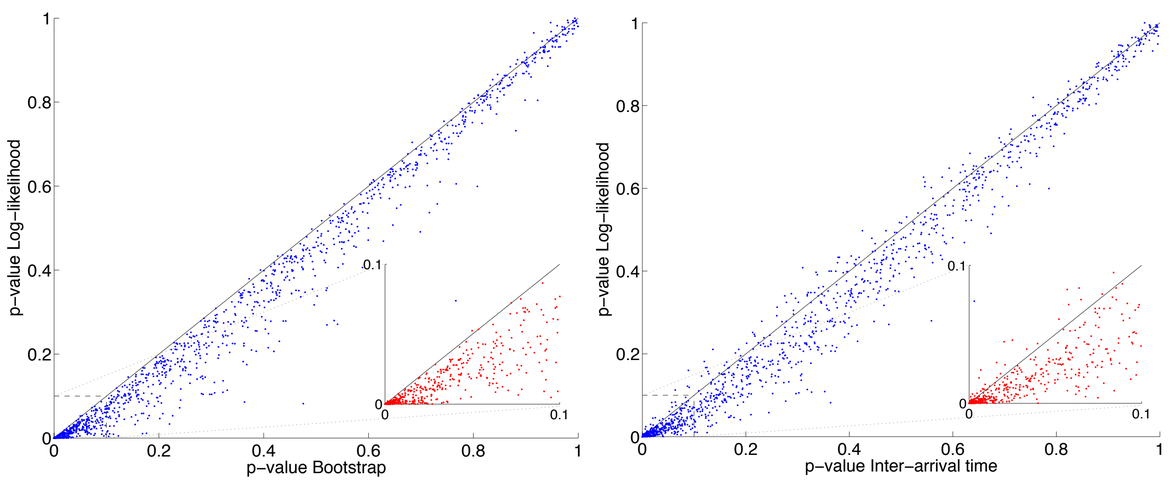

Figure 2: Comparison of p-values for the null hypothesis that the word is equally frequent in the two periods (1600-1639 and 1640-1681) of the Parsed Corpus of Early English Correspondence, for all words in the corpus. Both the bootstrap and inter-arrival time tests are often much more conservative than the log-likelihood ratio test.

Learning complex queries

- Linguists would often like to query a corpus for complex constructs
  - For example, *premodifying -ing participles* [4]
  - These are -ing participles that modify a noun, e.g., 'the *barking* dog'
- Straightforward queries have low recall because parsers and part-of-speech taggers are imperfect
- A query is essentially the same as a binary classifier
- We can *learn* complex queries just like training a classifier

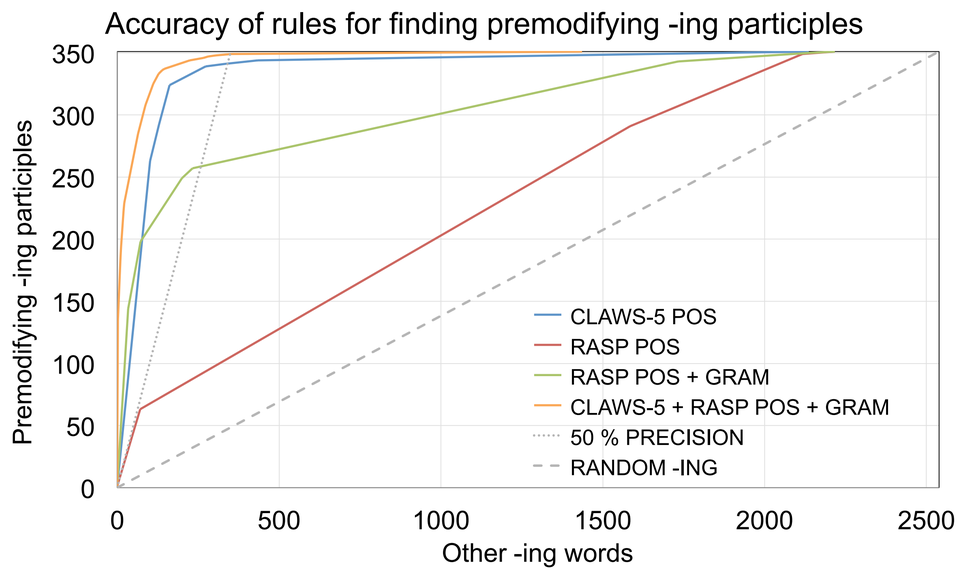

Figure 4: Precision and recall for classifiers based on several sources of information, based on a sample of 2902 *-ing* words, of which 351 are premodifying *-ing* participles, from the British National Corpus. Figure taken from [4].

References

[1] Altmann, Pierrehumbert & Motter 2009. "Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words". *PLoS One*, 4 (11): e7678.

[2] Katz 1996. "Distribution of content words and phrases in text and language modelling". *Natural Language Engineering*, 2 (1): 15–59.

[3] Lijffijt, Papapetrou, Puolamäki & Mannila 2011. "Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping". In *Proceedings of the ECML-PKDD 2011*.

[4] Vartiainen & Lijffijt 2012. "Premodifying -ing participles in the parsed BNC". In *Corpus Linguistics and Variation in English: Theory and Description.* Amsterdam/New York: Rodopi.