Literary Authorship Attribution with Phrase-Structure Fragments

Institute for Logic, Language and Computation
University of Amsterdam

Andreas.van.Cranenburgh@huygens.knaw.nl

Problem

- Stylometry is often done with superficial features, e.g. part-of-speech tags, certain function words.
- The result is a statistical decision that is hard to interpret.
- Can we find higher-level patterns in texts, e.g., in their syntactic structures?
- Can the resulting patterns characterize an author's style?

Contributions

- Stylometry based on full parse trees is viable
- . . . in this work applied to authorship attribution for evaluation purposes.
- Content words perform well when arbitrary fragments are considered.
- A mere 20 sentences w/phrase-structures already works for classification.
- A combination of trigrams and fragments performs better than either on its own.

Results

leave one out cross-validation for each work.
known author corpora: 15,000 sentences.

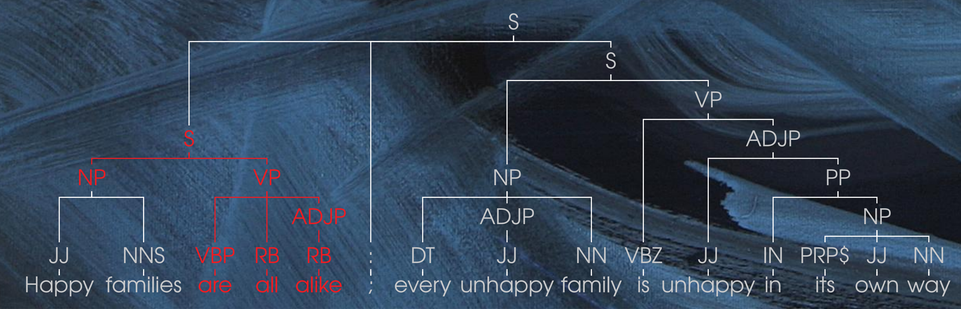
	20 test sentences			100 test sentences		
	trigrams	fragments	combined	trigrams	fragments	combined
Conrad	83.00	87.00	94.00	100.00	100.00	100.00
Hemingway	77.00	52.00	81.00	100.00	100.00	100.00
Huxley	86.32	75.79	86.32	89.47	78.95	89.47
Salinger	93.00	86.00	94.00	100.00	100.00	100.00
Tolstoy	77.00	80.00	90.00	95.00	100.00	100.00
average:	83.23	76.16	**89.09**	96.97	95.96	**97.98**

Federalist papers: **14** of 15 disputed/co-authored papers classified correctly.

Fragments

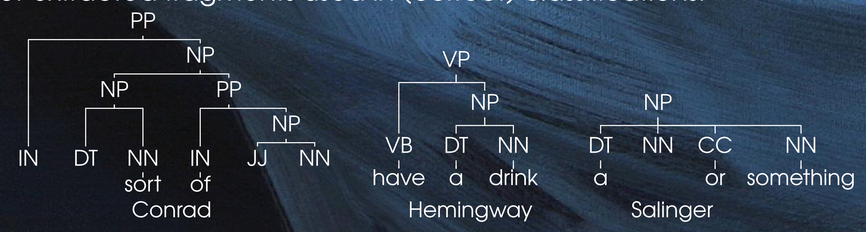
Notion of fragments (subset trees) taken from Data-Oriented Parsing (Scha, 1990; Bod, 1992):

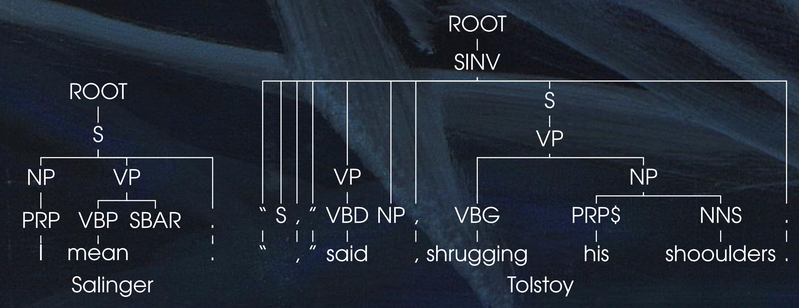
Definition. *A fragment f of a tree T is a connected subset of nodes from T, with $|f| \geq 2$, such that each node of f has either all or none of the children of the corresponding node in T.*



Examples

Examples of extracted fragments used in (correct) classifications:





Method

- Texts are parsed with an off-the-shelf parser (Stanford parser)
- Given two texts, their common maximal fragments are extracted with a tree-kernel method. This can be done in linear average time (Moschitti, 2006).
- Similarity of two texts defined as:

$$f(A, B) = \sum_{x \in A \sqcap B} \text{content\_words}(x)$$

where . . .

  - A and B contain phrase-structures of the sentences in a text.
  - the operation $A \sqcap B$ gives the maximal common fragments in A and B
  - content_words(x) for a fragment x gives the number of content words (adjective, adverb, noun, verb).

- To guess the author of a work:

$$\arg\max_{A \in \text{Authors}} \frac{f(A, B)}{1/|A| \sum_{t \in A} |t|}$$

Scores are normalized by dividing with the average number of nodes in the known texts of an author.

- As a baseline we use a similar model with trigrams instead of fragments.

Data

5 authors, 23 English texts:

Author (sentences)	Works (year of first publication)
Conrad, Joseph (25,889)	Heart of Darkness (1899), Lord Jim (1900), Nostromo (1904), The Secret Agent (1907)
Hemingway, Ernest (40,818)	A Farewell To Arms (1929), For Whom the Bell Tolls (1940), The Garden of Eden (1986), The Sun Also Rises (1926)
Huxley, Aldous (23,954)	Ape and Essence (1948), Brave New World (1932), Brave New World Revisited (1958), Crome Yellow (1921), Island (1962), The Doors of Perception (1954), The Gioconda Smile (1922)
Salinger, J.D. (26,006)	Franny & Zooey (1961), Nine Stories (1953), The Catcher in the Rye (1951), Short stories (1940–1965)
Tolstoy, Leo (66,237)	Anna Karenina (1877); transl. Constance Garnett, Resurrection (1899); transl. Louise Maude, The Kreutzer Sonata and Other Stories (1889); transl. Benjamin R. Tucker, War and Peace (1869); transl. Aylmer Maude & Louise Maude

Note that the works by Tolstoy are English translations from project Gutenberg; the translations are contemporaneous with the works of Conrad.

References

Rens Bod. 1992. A computational model of language performance: Data-oriented parsing. In *Proc. COLING*, pages 855–859.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *Proc. EACL*, pages 113–120.

Remko Scha. 1990. Language theory and language technology; competence and performance. URL `http://iaaa.nl/rs/LeerdamE.html`.

Part of project 'The Riddle of Literary Quality'
`http://literaryquality.huygens.knaw.nl`

The code used in the experiments is available at:
`http://github.com/andreasvc/authident`.

Painting: Christine Bittremieux (2007), Untitled. 70 × 100 cm. Oil on canvas. `www.bittremieux.nl`