Identification of Configurational Features for Authorship Attribution by Intrinsic Evaluation

Jussi Karlgren and Gunnar Eriksson

PAN 07 / SIGIR / Amsterdam

Rules, Constraints, and Conventions

Rules of language use operate on many levels and with varying force

Rule	Language	*Syntax, morphology*
Convention	Genre	*Lexical patterns, patterns of argumentation, tropes*
Free	Author	*Repetition, organisation, elaboration*

The Dynamic Nature of Language

A non-contentious claim:

Individual variation →
→ Conventionalisation →
→ Grammaticalisation

Textuality and convention

Rules and conventions tend to operate locally:

- *text scope is wide; reader view is narrow*
- *too many degrees of freedom for convenient rule formulation.*

Features, Measurement, and Aggregation

Features
: are observable linguistic or textual items: words, constructions, etc.

Measurement
: of features can be observed occurrence, frequency of occurrence, variation, etc.

Aggregation
: of measurements can be averages of various sorts (mean, mode), configurations of sequences of occurrence, etc.

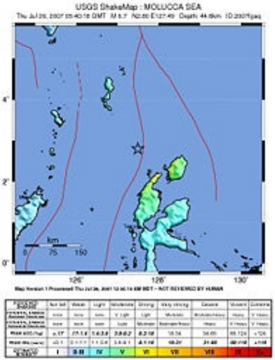
Pointwise vs. Configurational

Our claims:

Features for *authorship analysis* should be selected from areas where conventional and grammatical forces are weak;

Aggregation of features is better done using *sequence* rather than averaging *pointwise* non-contextual observations.

Example: event



Example: Text 1

A powerful earthquake jolted eastern Indonesian islands in North Maluku province Thursday, prompting government authorities to a tsunami warning. The quake, measuring 6.6 on the Richter scale, took place at about 0540 GMT, shaking Halmahera and nearby islands in North Maluku province, said Fauzi, an official at Jakarta's Meteorology and Geophysics Agency. According to the US Geological Survey USGS, the quake was measured at 7.0 on the Richter scale. "We have issued a warning that the quake could potentially trigger a tsunami," Fauzi told Deutsche Presse-Agentur dpa. He said the quake took place about 57 kilometres beneath the seabed. No immediate casualties or injuries were reported from the quake. Indonesia is located in the Pacific volcanic belt known as the "Ring of Fire," where earthquakes and volcanoes are common. On December 26, 2004, a massive 9.0-magnitude earthquake, which triggered gigantic tidal waves, devastated thousands of homes and buildings along the coastline of northern Sumatra, leaving around 170,000 people dead or missing in Indonesia and thousands more dead and injured along the Indian Ocean coastline.

Example: Text 2

A powerful earthquake rocked eastern Indonesia on Thursday, sending residents fleeing from swaying homes and hospitals, authorities and witnesses said. There were no immediate reports of damage. The quake, which had a preliminary magnitude of 7, triggered a tsunami warning but the alert was quickly lifted after it became clear no destructive waves had been generated, the country's geophysics agency said. The earthquake struck under the Maluku Sea at a depth of 20 miles, the U.S. Geological Survey said on its Web site. The quake's epicenter was more than 130 miles north of Ternate city. "We felt a strong tremor for almost a minute, people ran in panic from buildings, said George Rajaloa, a resident in Ternate. "Children are crying and their mothers are screaming, but there is no damage in my area." Indonesia, the world's largest archipelago, is prone to seismic upheaval due to its location on the so-called Pacific "Ring of Fire," an arc of volcanoes and fault lines encircling the Pacific Basin. In December 2004, a massive earthquake struck off Sumatra island and triggered a tsunami that killed more than 230,000 people in a dozen countries, including 160,000 people in Indonesia's westernmost province of Aceh. Just over a year ago, another quake-generated tsunami killed around 600 people on Java island.

Example: Text 3

According to the United States Geological Survey USGS a strong magnitude 6.9 earthquake has struck Indonesia in the Molucca Sea approximately 220 kilometers 135 miles north of Ternate, Maluku Islands, Indonesia at a depth of 44.6 kilometers 27.7 miles. The Japan Meteorological Agency reports the quake at a magnitude 7.0 with a depth of 50 km. An unnamed official with the USGS says "there is a potential that a tsunami might develop, judging from the magnitude," but no tsunamis were reported. "We have lifted the warning. After monitoring, there were no signs of tsunami," said the Indonesian head of the country's geology agency, Fauzi.Initially, Fauzi issued a tsunami warning saying "we have issued a warning that the quake could potentially trigger a tsunami."There are no reports of injuries, deaths or damage. One resident in Ternate said that he "felt a strong tremor for almost a minute, people ran in panic from buildings. Children are crying and their mothers are screaming but there is no damage in my area." Earlier the National Oceanic and Atmospheric Administration NOAA had issued a tsunami bulletin stating that local high waves could be possible, but a widespread tsunami is "not expected based on historical earthquake data."

Example: Features

	Text 1		*Text 2*		*Text 3*	
Sentences	8		10		10	
Words	175		213		203	
wps	6.6		6.2		6.2	
cpw	21.9		21.3		20.3	
clause	4		6		5	
adv	4		6		4	
1	-	+	+	+	-	+
2	+	+	-	-	-	-
3	-	+	+	-	+	-
4	+	-	+	+	-	-
5	+	-	-	-	+	-
6	-	-	+	+	+	+
7	-	-	+	-	-	-
8	+	+	-	+	+	+
9			+	+	+	-
10			-	+	+	+

Sub-genres of Glasgow Herald

ARTICLETYPE	*n*
advertising	522
book	585
correspondence	3659
feature	8867
leader	681
obituary	420
profile	854
review	1879
total	17467

Authors

Select the 244 authors with $>$ 500 sentences in the corpus.

Feature: Adverbials

- Measure of topical elaboration
- The occurrence of more than one adverbial expression of any type in a sentence

On Sunday, an earthquake struck off the of coast Sumatra.

Feature: Clauses

- A measure of syntactic complexity
- The occurrence of more than two clauses of any type in a sentence

Children are crying and their mothers are screaming but there is no damage in my area.

Aggregation

We need an aggregation which preserves sequential order information.

Average occurrence frequencies will not do this.

Computing transitions from one observation of a feature to the next is a candidate methodology.

Transition Patterns

Feature space for varying window sizes

window size	patterns	number patterns
1	1, 0	2
2	11, 10, 01, 00	4
3	111, 110, 101, 100 011, 010, 001, 000	8
4	1111, ..., 0000	16
5	11111, ..., 11101, 11100, ..., ..., 00000	32

Probability Estimates

For each setting, we obtain an estimate of probabilities for observing some given sequence of observations:

$$p_3(correspondence) =$$

$$= \{p_{111}, p_{110}, p_{101}, p_{100}, p_{011}, p_{010}, p_{001}, p_{000}\} =$$

$$= \{0.0069, 0.0654, 0.00903, 0.155, 0.00454, 0.0363, 0.0486, 0.674\}$$

Evaluation

Claim The quest for the optimal features and measures is better served by intrinsic than than extrinsic evaluation.

Kinds of evaluation

Extrinsic

- evaluation by task application
- guarantees validity
- introduces noise

Intrinsic

- evaluation by inspecting representation
- higher explanatory power
- modular with respect to task
- less risk of overfitting or programmer error

Intrinsic evaluation of features for author identification

Find a knowledge representation and features that give purchase to separation of test corpus categories.

Our candidate: Kullback-Leibler Divergence ($\approx$ Information Gain).

- Measures difference between two probability distributions.
- Here, a symmetric variation is used.
- We want to find a *large* difference between distributions.

Procedure

8 genres; 244 authors, with repeated sampling of eight authors from set.
The sum of all pairwise K-L divergence scores for the set of eight categories (genres or authors) is computed.

Results

K-L divergence sums for features CLAUSE and ADV,
window sizes 1 to 5,
comparison between author and genre categories.

Window size	GENRE		AUTHOR	
	CLAUSE	ADV	CLAUSE	ADV
1	0.5129	0.1816	0.7254	0.4033
2	0.8061	0.3061	1.3288	0.8083
3	1.1600	0.4461	2.1577	1.2211
4	1.4556	0.6067	2.3413	1.8111
5	1.7051	0.7650	3.0028	2.2253

1. Relative difference between AUTHOR and GENRE larger for larger window sizes.
2. ADV catches up with CLAUSE for larger window sizes.

Take home

- Configurational features – sequential aggregation of some observed item – make a difference.
- Author categories and genre categories can and should be identified differently: in the one case identifying conventions, in the other, avoiding them.
- Intrinsic evaluation of knowledge representation should be talked about more.

Choice points left by the wayside

- Is Kullback-Leibler divergence the right measure?
- Is summing pairwise divergences the best way of modelling the consistency of a set of category models?
- What better kernel features – beyond adverbial and clause count – should we use?
- How can we *combine* features – preferrably without resorting to weighted linear combinations?