Unsupervised Detection of Anomalous Text

Pavel Veselý

Source

- „Unsupervised Detection of Anomalous Text“ by David Guthrie
  - PHD thesis at University of Sheffield, 2008

Dictionary (TODO Change)

- Anomaly - a deviation from the common rule, type, arrangement, or form
- Texts – documents or segments of documents (in this thesis segments of 100, 500 and 1000 words)
- Unsupervised detection -

Motivation

- Detecting plagiarism without access to source text
  - Koppel, Seidman – Automatically Identifying Pseudepigraphic Texts
  - Klára Kufová – Anomaly Detection in Text
- Getting more homogeneous text set
  - Guthrie, Guthrie, Wilks – An Unsupervised Approach for the Detection of Outliers in Corpora
- Out of topic posts in forums

Process

1. Represent texts as numerical vectors
2. Measure distance of vectors from rest of set
3. Set threshold for anomaly

Text representation

- Numerical vector of 166 features
  1. Simple Surface Features (19)
  2. Readability Measures (7)
  3. Obscurity of Vocabulary Features (7)
  4. Part of Speech and Syntax Features (11)
  5. Rank Features (8)
  6. Emotional Tone Features (114)

Simple Surface Features I

1. Average sentence length
2. Average word length
3. Average number of syllables per word
4. Percentage of all words that have 3 or more syllables
5. Percentage of all words that only have 1 syllable
6. Percentage of long sentences (sentences greater than 15 words)
7. Percentage of short sentences (sentences less than 8 words)
8. Percentage of sentences that are questions
9. Percentage of all characters that are punctuation characters
10. Percentage of all characters that are semicolons

Simple Surface Features II

11. Percentage of all characters that are commas
12. Percentage of all words that have 6 or more letters
13. Percentage of word types divided by the number of word tokens
14. Percentage of words that are subordinating conjunctions (then, until, while,since, etc.)
15. Percentage of words that are coordinating conjunctions (but, so, but, or, etc.)
16. Percentage of sentences that begin with a subordinating or coordinating conjunctions
17. Percentage of words that are articles
18. Percentage of words that are prepositions
19. Percentage of words that are pronouns

Readability Measures

1. Flesch-Kincaid Reading Ease
2. Flesch-Kincaid Grade Level
3. Gunning-Fog Index
4. Coleman-Liau Formula
5.Automated Readability Index
6.Lix Formula
7. SMOG Index

Obscurity of Vocabulary Usage

- Number of words in text, that have relative frequency in corpus (Gigaword) as follows:
  1. Top 1000 words
  2. Top 5000 words
  3. Top 10,000 words
  4. Top 50,000 words
  5. Top 100,000 words
  6. Top 200,000 words
  7. Top 300,000 words

Part of Speech and Syntax Features

1.Percentage of words that are adjectives

2.Percentage of words that are adverbs

3. Percentage of words that are interrogative words (who, what, where when, etc.)

4. Percentage of words that are nouns

5.Percentage of words that are verbs

6.Ratio of number of adjectives to nouns

7.Percentage of words that are proper nouns

8.Percentage of words that are numbers (i.e. cardinal, ordinal, nouns such as dozen, thousands, etc.)

9.Diversity of POS tri-grams

$$POS\,Trigram\ Diversity = \left(\frac{number\ of\ different\ POS\,trigrams}{total\,number\ of\ POS\,trigrams}\right) x\,100$$

Rank Features

1. Distribution of POS tri-grams list
2. Distribution of POS bi-gram list
3. Distribution of POS list
4. Distribution of Articles list
5. Distribution of Prepositions list
6. Distribution of Conjunctions list
7. Distribution of Pronouns list
8. Distribution of Adverbs list

General Inquirer Dictionary

- Capturing sentiment of text
- Quantify the connotative meaning of isolated words
- 13,000 root words mapped into 114 categories
  - most words assigned to more than one category
  - The two largest categories are ‘positive’ (1,915 words) and ‘negative’ (2,291 words)

Measuring the distances

- ClustDist
  - A distance based on average linkage clustering
- **SDEDist**
  - The Stahel-Donoho Estimator distance
- Pcout
  - The weights calculated by the PCout algorthim
- MeanComp
  - Distance from the mean of all other segments in the data
- **TxtCompDist**
  - Method developed by authors that uses the distance from the textual complement

SDEDist

- Projection can give an scalar distance to the center of all observations
- Find the direction that, when used for projection, gives maximum distance

$$SDEDist\left(\vec{x}, V\right) = \max_{\vec{a}} \frac{\overrightarrow{x^T}\vec{a} - median\left(\boldsymbol{V}\vec{a}\right)}{mad\left(\boldsymbol{V}\vec{a}\right)}$$

- Infinitely many directions – problem of finding good direction set

TxtCompDist

- Distance from textual complement (the union of the remaining texts)

$$TxtCompDist(\vec{x}, \mathbf{V}) = d(\vec{x}, \vec{c_x})$$

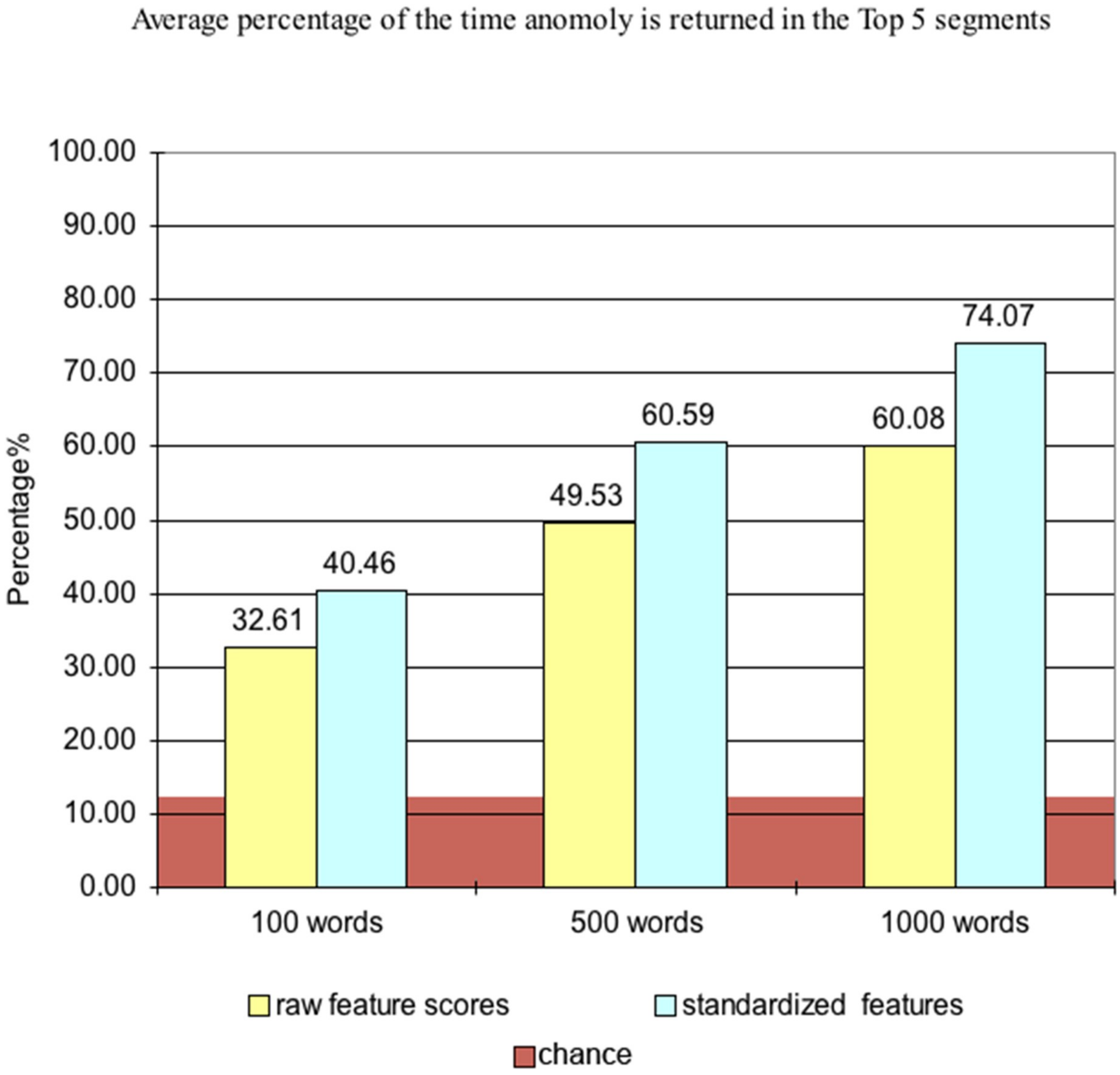
- Designed by authors
- Better use of features requiring larger texts (POS trigrams, adverb preference)

Experiment Data

- Artificial data set
  - Documents of 51 segments, 1 of which is anomalous
  - Created as random bag of segments from 2 sources (50 segments by 1)
- Segments of length 100, 500 and 1000 words

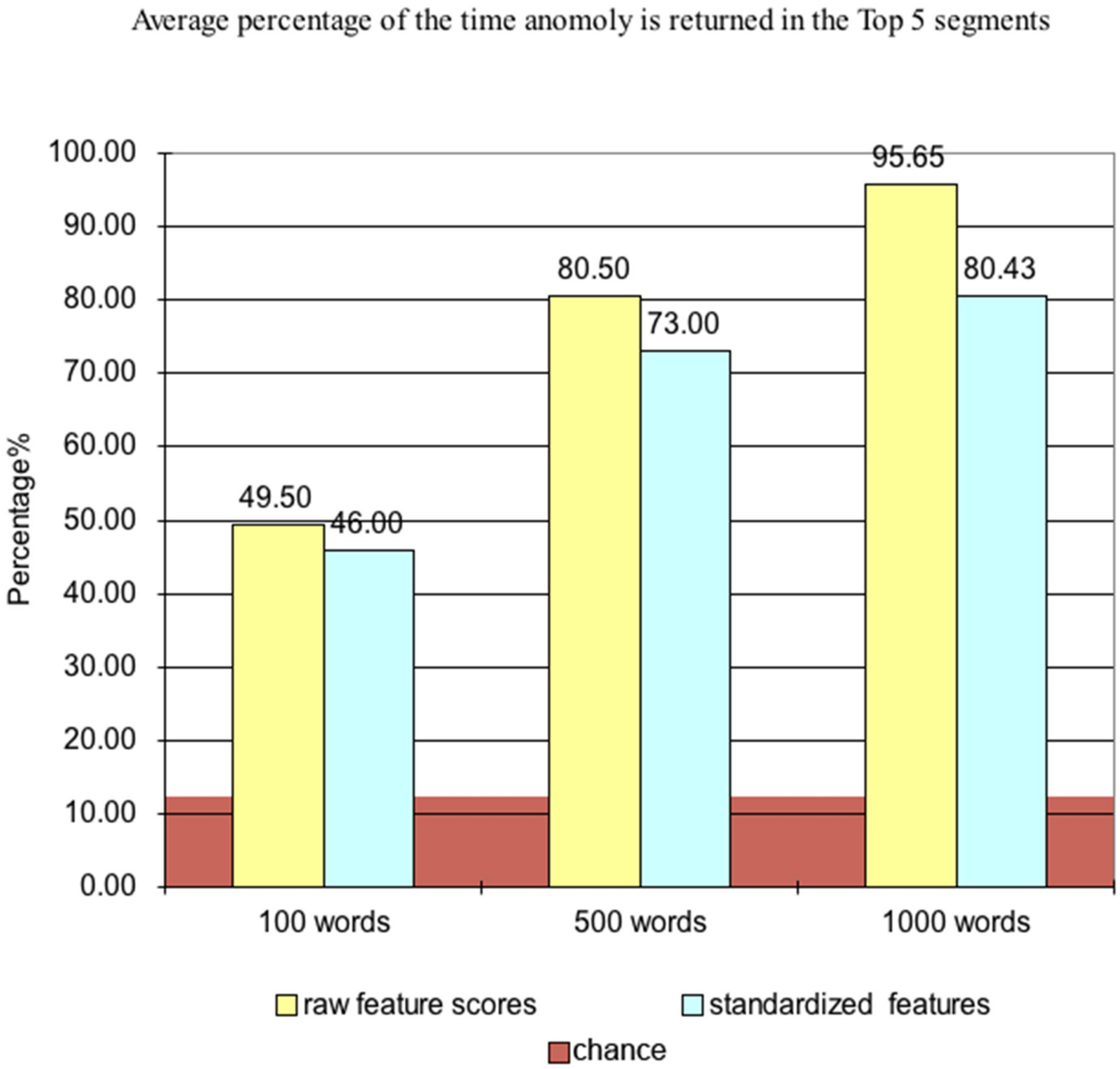
Authorship Tests

- 8 Victorian authors



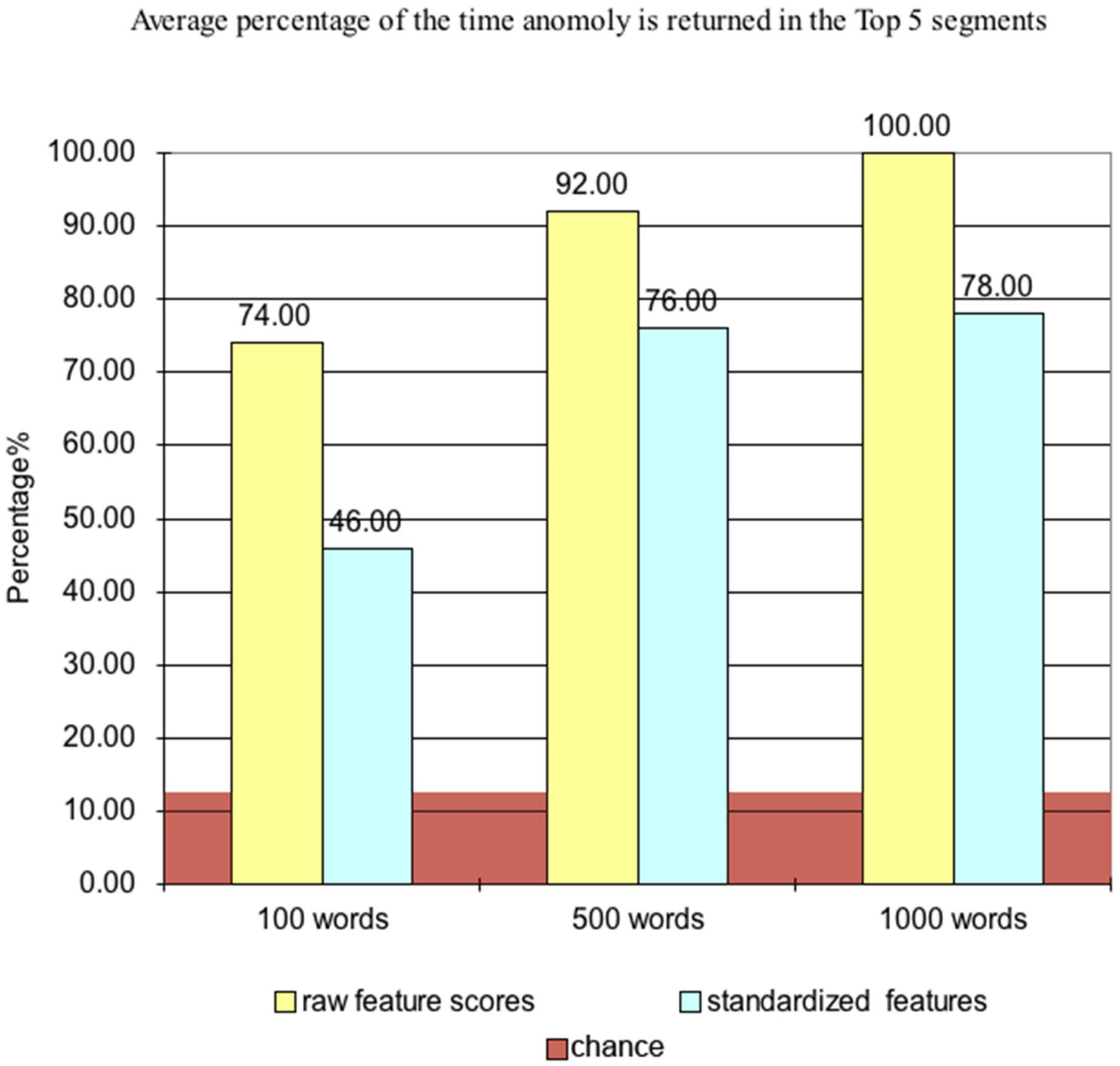
Fact versus Opinion

- Factual news versus editorials



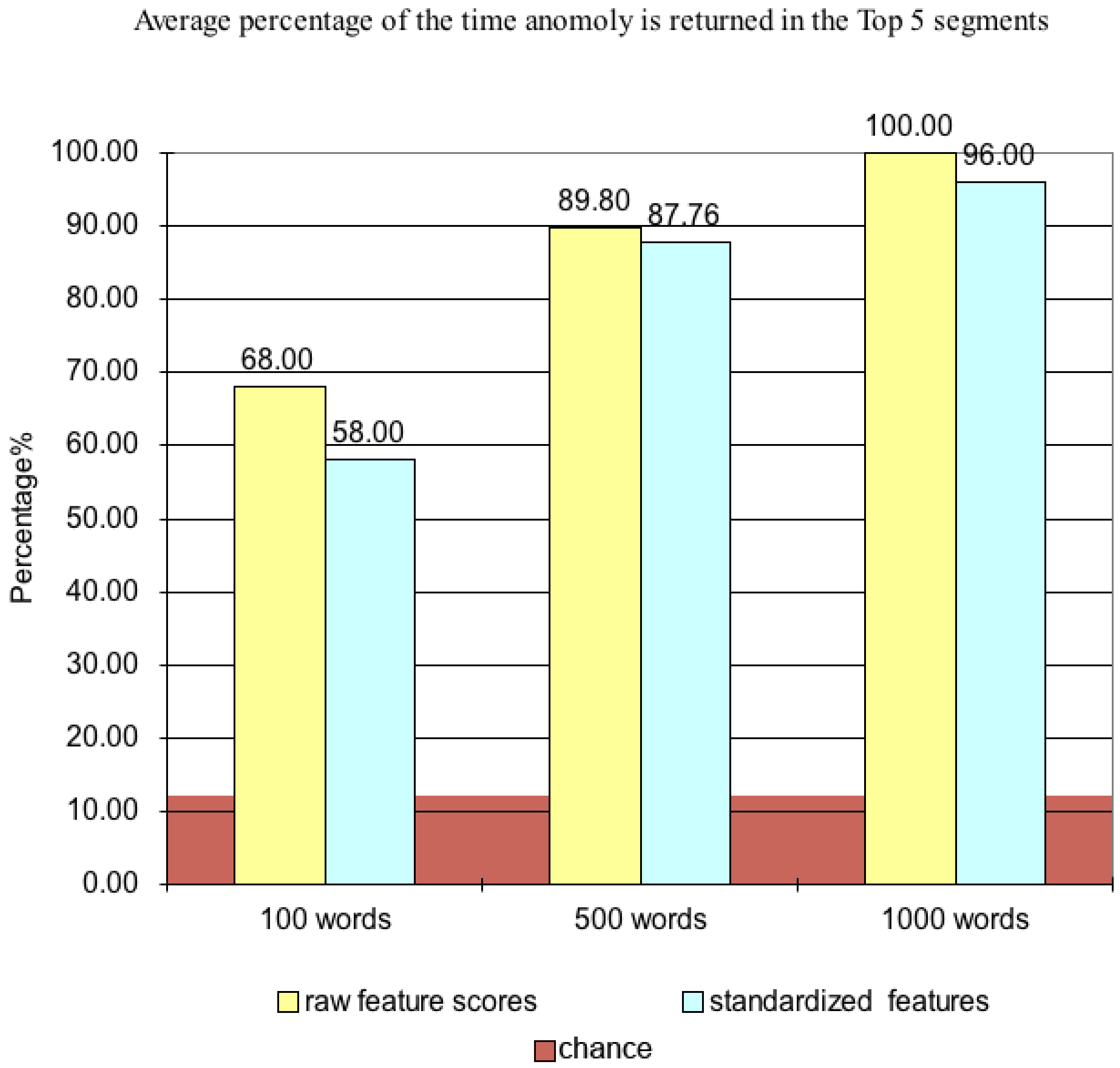
Genre Difference

- Newswire versus Anarchist Cookbook



Machine Translation

- English Newswire versus Chinese newswire translated by 2008 Google Translate

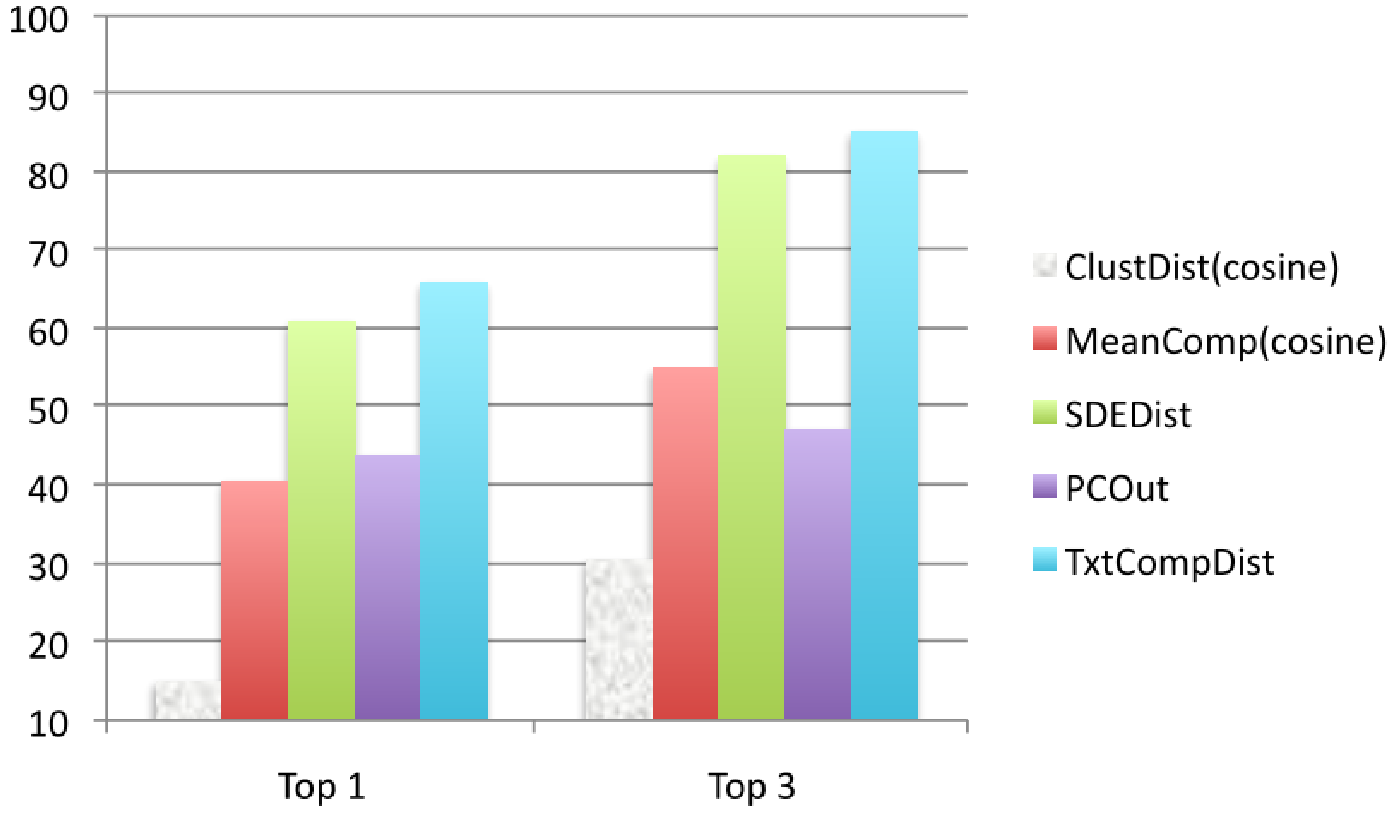


Experiment Conclusion I

- Finding anomaly as top 1
  - Random chance – 2%
  - Average on 100 words – 32%
  - Average on 1000 words – 68%
- Best on Google Translate (2008) – 96%
- Works very well on anomalies different in style or genre

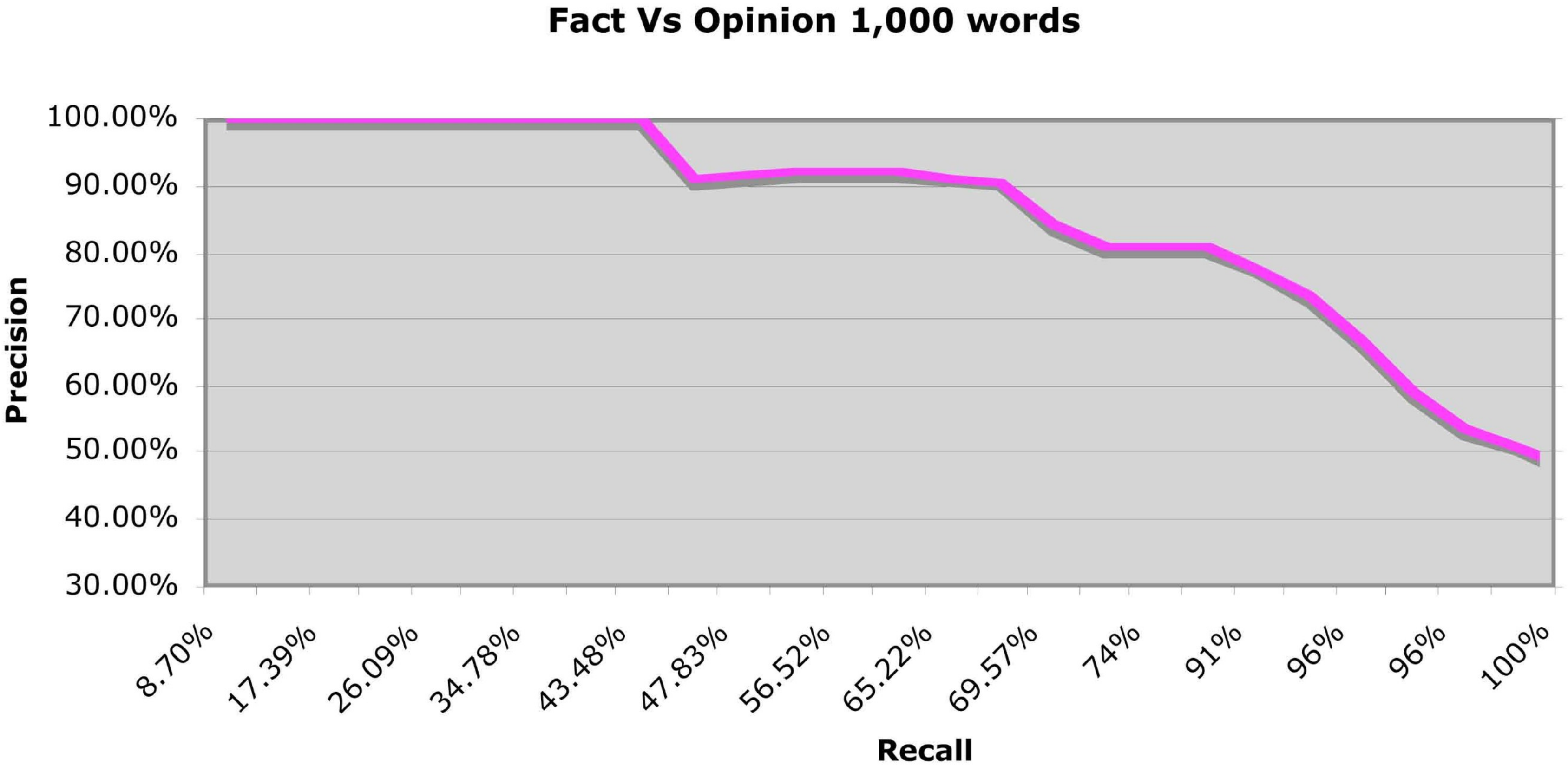
Experiment Conclusions II

- Best metric – TxtCompDist
- Stahel-Donoho based method close second



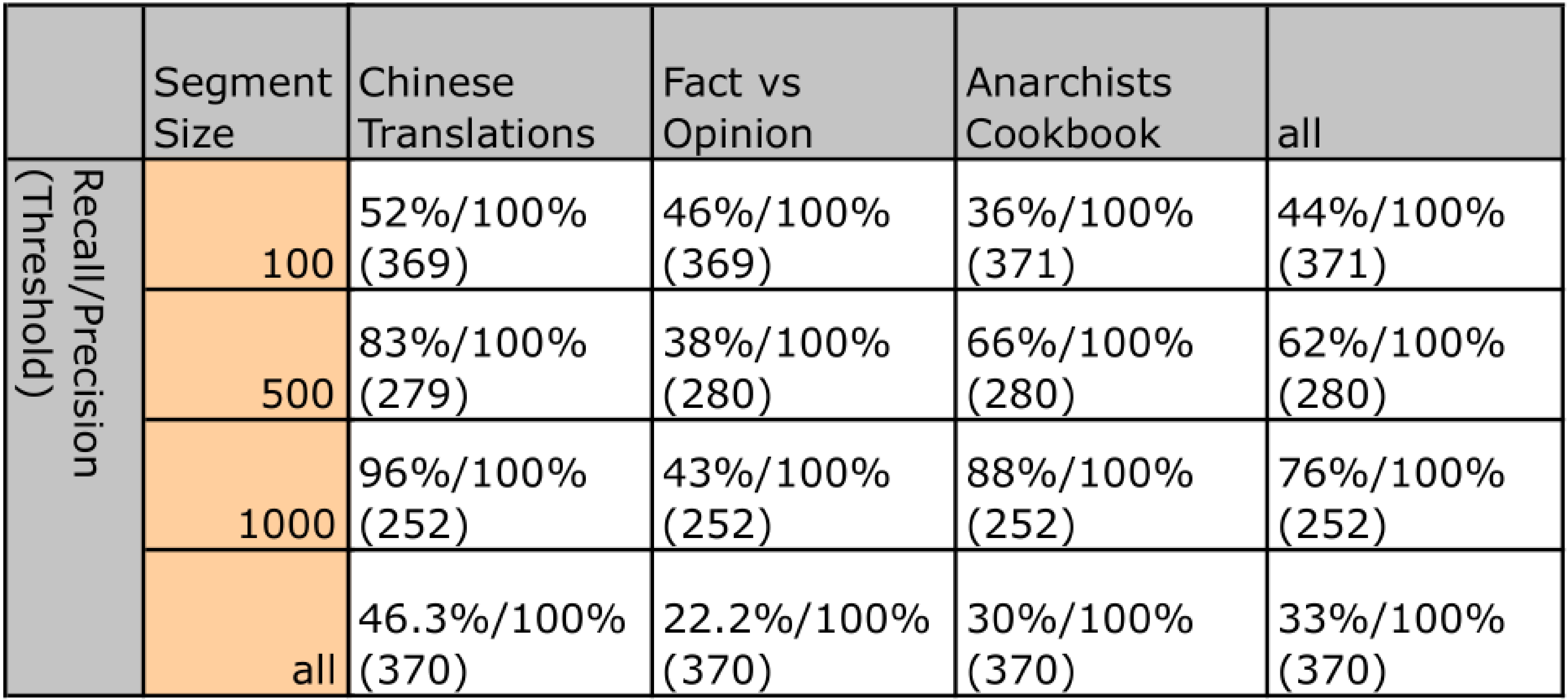
Precision and Recall

- Setting threshold of anomaly score
- Compromise between precision and recall



Thresholding

- Fixed 100% precision, maximizing recall



	Segment Size	Chinese Translations	Fact vs Opinion	Anarchists Cookbook	all
Recall/Precision (Threshold)	100	52%/100% (369)	46%/100% (369)	36%/100% (371)	44%/100% (371)
	500	83%/100% (279)	38%/100% (280)	66%/100% (280)	62%/100% (280)
	1000	96%/100% (252)	43%/100% (252)	88%/100% (252)	76%/100% (252)
	all	46.3%/100% (370)	22.2%/100% (370)	30%/100% (370)	33%/100% (370)

Feature Selection

- Based on ability to differentiate anomalies
- Best features are the same over all experiments
  1. Gunning-Fog Index
  2. Number of passive sentences
  3. Flesch-Kincaid Reading Ease
  4. Percenteage of sentences over 15 words
- Worst features differ, but mostly sentiment based

Real Data Experiments

- TODO
- Klára
- Guthrie – corpora

Summary

- TODO