

The Computer and Natural Language (Ling 445/515)

Topic 3.1: Text Classification

Markus Dickinson
Dept. of Linguistics, Indiana
Autumn 2008

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

1 / 24

Authorship Attribution

- ▶ **Authorship attribution** is the process of identifying who wrote a text.
- ▶ Potential applications include:
 - ▶ Author Identification (Madison or Hamilton ... who penned *The Federalist Papers*?)
 - ▶ Forensic Evidence (suicide or murder ... who wrote the note?)
 - ▶ Plagiarism Detection (pass or fail ... who did the work?)

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

2 / 24

Text Classification

- ▶ Authorship attribution is a **text classification** task.
- ▶ **Text classification** = take documents and a set of relevant categories and figure out which documents belong into which category.
 - ▶ For example, email sent to the New York Times could be classified into letters to the editor, new subscription requests, complaints about undelivered papers, job inquiries, proposals to buy ad pages, and others
- ▶ Other related classification tasks we'll talk about:
 - ▶ Language Identification (English or French ... what language is this anyway?)
 - ▶ Email Filtering (Spam or Ham ... who is sending you email?)
- ▶ Can we do such classification tasks automatically?

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

3 / 24

Language identification

- ▶ We can attempt to classify documents according to the language a document is (mostly) written in.
- ▶ Can sometimes tell by
 - ▶ which characters are used,
 - ▶ e.g. *Liebe Grüße* uses ü and ß → German
 - ▶ which character encoding is being used
 - ▶ e.g., ISO 8859-8 is used to encode Hebrew characters → text is written in Hebrew
- ▶ But how can you tell if you are reading English vs. Japanese transliterated into the Roman alphabet? Or Swedish vs. Norwegian? And all phonetically transcribed text is encoded in the same IPA encoding!
- ▶ Consider what you base your guess on when I ask whether the following is Portuguese or Polish:
Czy brak planów zagospodarowania hamuje rozwój Warszawy?

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

4 / 24

Language identification

N-grams

- ▶ One simple technique for identifying languages is to use **n-grams** = stretch of n tokens (i.e., letters or words):
 - ▶ Go through texts for which we know which language they are written in and store the n-grams of letters found, for a certain n .
 - ▶ e.g., extracting the trigrams (3-grams) for the last sentence we'd get: *Go , o t, th, thr, hro, rou, ...*
 - ▶ This provides us with an indication of what sequences of letters are possible in a given language (and how frequent they occur).
 - ▶ e.g., *thr* is not a likely Japanese string.
- ▶ How do we make this more concrete?

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

5 / 24

Language identification

Frequency distributions

- ▶ Store a **frequency distribution** of trigrams, i.e., how many times each n-gram appears for a given language.

n-gram	English	Japanese
aba	12	54
ace	95	10
act	45	1
arc	8	0
...

- ▶ Now, apply the frequency distribution to a new text and use it to help calculate the probability of the text being a particular language.
 - ▶ Compare each n-gram to see if it is more likely to be English or Japanese.
 - ▶ See which language won the most comparisons.

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

6 / 24

Language identification

Different techniques

- ▶ Although n-grams do not capture abstract linguistic knowledge, they are a simple and surprisingly effective technique, used throughout computational linguistics.
- ▶ Another simple technique for language identification would be to look for keywords in the documents, e.g., *capture* → English, *je* → French, etc.
 - ▶ Requires knowledge which words are the best indicators for a particular language.
 - ▶ Words occurring frequently and independent of the topic of the text are best, e.g., so-called function words like articles (e.g., in English *the, a, . . .*), complementizers (e.g., in English *that, whether, if, . . .*).

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

7 / 24

Identifying the Author

- ▶ In a classic study, Mosteller and Wallace (1964) applied authorship detection techniques to *The Federalist Papers*.
- ▶ *The Federalist Papers* were a series of 85 articles written between 1787 and 1788 by James Madison, Alexander Hamilton and John Jay to persuade New York to ratify the Constitution.
- ▶ Some of the papers were clearly written by one of the three; 12 are in question, written either by Hamilton or Madison.
- ▶ Mosteller and Wallace examined the frequency of various words in the disputed papers and compared each to a model of known Hamilton writings and known Madison writings.

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

8 / 24

Stylometry

- ▶ **Stylometry** defines the features of an author's style and measures those features in two or more texts to determine the similarity between the texts.
- ▶ The more similar the styles, the more likely two texts are to be written by the same author.
- ▶ The idea is that style operates at a subconscious level, which makes it more consistent (and perhaps measurable?).
- ▶ In other words, writing style is a "linguistic fingerprint."

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

9 / 24

Stylometric Approach

- ▶ The basic approach:
 - ▶ Extract style markers
 - ▶ Use the markers to classify texts
- ▶ Style markers may be based on words, grammar or a combination.

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

10 / 24

Lexical Style Markers

- ▶ **Lexical style markers** are words that give clues about authorship.
- ▶ There are two types of markers: vocabulary richness and frequency of function words.
 - ▶ **Function words** such as "to" and "that" carry little meaning but occur often in a
 - ▶ Function words are independent of topic, but the idea is that *which* function words you choose and *where* you use them are enough to identify you as an author.
- ▶ How can we use lexical markers to detect plagiarism?

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

11 / 24

Frequency of Function Words

- ▶ An example of two authors' use of function words, gathered from AP news stories by Zhao and Zobel (2005).

	a	and	for	in	is	of	that	the
Barry Schweid	6.28	9.22	4.94	6.50	1.62	14.66	1.89	29.13
Don Kendall	9.75	7.08	2.36	7.99	3.05	13.16	5.73	41.29

- ▶ The Signature Text Analysis program (<http://www.etext.leeds.ac.uk/signature/>) is designed to help you determine such stylometric indicators

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification
Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words
Plagiarism Detection
What is plagiarism?
Plagiarism Example
Plagiarism Detection
Detection Goals
Previous Approaches
References

12 / 24

What is plagiarism?

- ▶ Clough (2003) defines **text reuse** is the deliberate or unintentional use of existing text for the creation of a new text.
 - ▶ **Plagiarism** is one kind of text reuse.
 - ▶ Reusing newswire text in journalistic publications is another instance of text reuse.

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

13 / 24

Types of Plagiarism

- Clough (2003) outlines six forms of plagiarism:
1. Word-for-word – Whole phrases, sentences or passages are copied, but not attributed.
 2. Paraphrasing – The unattributed source material is rewritten, but is still recognizable in the new text.
 3. Secondary Source – Sources are cited, but extracted from a secondary source (not the original).
 4. Source Form – A source’s argument structure/text organization is copied.
 5. Ideas – Thoughts (independent of form) are copied without attribution.
 6. Authorship – Authorship of an entire text is falsely claimed.

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

14 / 24

Word-for-Word Plagiarism: Source*

The Passage as It Appears in the Source:
Critical care nurses function in a hierarchy of roles. In this open heart surgery unit, the nurse manager hires and fires the nursing personnel. The nurse manager does not directly care for patients but follows the progress of unusual or long-term patients. On each shift a nurse assumes the role of resource nurse. This person oversees the hour-by-hour functioning of the unit as a whole, such as considering expected admissions and discharges of patients, ascertaining that beds are available for patients in the operating room, and covering sick calls. ... (Chase, 1995, p. 156)

*Example From the Writing Center at University of Wisconsin-Madison
(http://www.wisc.edu/writing/Handbook/QPA_paraphrase.html).

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

15 / 24

Word-for-Word Plagiarism: Copy

Critical care nurses have a hierarchy of roles. The nurse manager hires and fires nurses. S/he does not directly care for patients but does follow unusual or long-term cases. On each shift a resource nurse attends to the functioning of the unit as a whole, such as making sure beds are available in the operating room, and also has a patient assignment. ...

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

16 / 24

Recognizing Plagiarism (1)

- The following factors may indicate plagiarism:
- ▶ Vocabulary use beyond the skill level of the writer (Ex: technical/advanced terms).
 - ▶ A drastic change in the quality of writing compared to previous submissions.
 - ▶ Style or vocabulary inconsistencies within a text.
 - ▶ Choppy text that lacks transitions or smooth flow, indicating a “cut-and-paste” job.

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

17 / 24

Recognizing Plagiarism (2)

- ▶ Significant similarity between multiple submissions.
- ▶ Similar errors between multiple submissions (Ex: the same spelling/grammar errors).
- ▶ References that appear in the text but not the bibliography.
- ▶ Lack of a consistent bibliographic style within the body or references section of text.

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

18 / 24

Plagiarism Detection

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection

Detection Goals
Previous Approaches

References

19 / 24

1. Detection in a single text:
 - ▶ Identify inconsistencies within a text
 - ▶ Find sources for the inconsistencies
2. Detection across multiple texts:
 - ▶ Identify unacceptable collaborations
 - ▶ Identify direct copying

Detection Goals

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection

Detection Goals
Previous Approaches

References

20 / 24

- We would like to . . .
- ▶ Maximize **true positives** (texts *correctly* marked as instances of plagiarism) and **true negatives** (texts *correctly* marked as not instances of plagiarism).
 - ▶ Minimize **false positives** (texts *incorrectly* marked as instances of plagiarism) and **false negatives** (texts *incorrectly* marked as not instances of plagiarism).

Previous Approaches

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

21 / 24

- ▶ Most work on plagiarism detection has been for identifying instances across texts, rather than within a single text (Clough, 2003).
- ▶ As the plagiarism becomes subtler, the task becomes harder. (i.e., It is easier to catch direct copying.)
- ▶ Automatic detection focuses on finding suspicious features – indicators of plagiarism.

Indicators of plagiarism

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

22 / 24

- ▶ The idea is that the more similar two texts are, the more likely it is that one of the text is derived (plagiarized) from the other.
- ▶ Possible indicators include vocabulary use, word length, syllable structure, rhyme and grammar
 - ▶ Q: what features or methods would you use to detect the similarities/differences between 2 texts?
- ▶ Indicators are used to flag texts for later human inspection.

Effective Detection

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

23 / 24

Effective detection requires:

1. Reliable, measurable indicators of plagiarism
2. A method for comparing indicators
3. A method for calculating similarity between texts

You can see a comparison of modern systems at:
http://www.jiscpas.ac.uk/documents/resources/PDReview-Reportv1_5.pdf

And a journal of all things related to plagiarism:
<http://www.plagiary.org/>

References

- These slides were developed using the following sources:
- ▶ Clough, Paul. 2003. *Old and new challenges in automatic plagiarism detection*. National Plagiarism Advisory Service.
 - ▶ Keselj, Vlado, Peng, Fuchun, Cercone, Nick and Thomas, Calvin. 2003. N-gram-based author profiles for authorship attribution. In *Proceeding of the Pacific Association for Computational Linguistics (PACLING'03)*, Dalhousie University, Halifax, Nova Scotia, Canada, August 2003.
 - ▶ Putnins, Talis, Signoriello, Domenic, Jain, Samant, Berryman, Matthew and Abbott, Derek. 2005. Advanced text authorship detection methods and their application to biblical texts. In *Proc. SPIE: Complex Systems 6039*. Brisbane, Qld., Australia, December 11-14, 2005.
 - ▶ Zhao, Ying and Zobel, Justin. 2005. Effective and scalable authorship attribution using function words. In *Proceedings of the AIRS Asia Information Retrieval Symposium*, Jeju Island, Korea, October 2005. pp. 174-189.

Computers and Language
Topic 3.1: Text Classification

Introduction
Text Classification
Language Identification

Authorship Attribution
Author Identification
Stylometry
Lexical Markers
Lexical Markers: Function Words

Plagiarism Detection
What is plagiarism?
Plagiarism Example

Plagiarism Detection
Detection Goals
Previous Approaches

References

24 / 24