Seminar on: Detecting Latent User Properties in Text

L715/B659
Dept. of Linguistics, Indiana University
Fall 2014

Goals of this class

Particular goals of this seminar:
- See the connections between what are at times disparate topics
- Understand the theoretical underpinnings of making such inferences
  - e.g., explore a range of linguistic features
- Obtain practice building a practical system

More generally, my goal in a seminar is to see you learn to:
- Develop as researchers
- Collaborate together on fun topics
⇒ Think of this more like a research lab than a class
  - I hate to quote High School Musical, but: “We’re all in this together.”

Specific topics

The specific topics will include:
- Authorship attribution
- Plagiarism detection
- Deception detection
- Author profiling (e.g., sex) & attribution in social media
- Dialect identification
- Native language identification
- Language proficiency identification

Part of that will depend upon student interest
- ... and maybe we’ll come up with more in the process?

Authorship attribution

Authorship attribution: who is the author of a document?
- Given a set of authors, which one (if any) wrote a particular work?
- Authorship verification: did a particular person write a particular work or not?
  - More challenging problem
  - Connects to forensic linguistics

http://www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-web/author-identification.html

Plagiarism detection

Plagiarism detection can be seen as a subpart of authorship attribution
- Is the author of the document who they claimed to be?

This can be divided into two tasks:
- Source retrieval: find source documents for a suspicious document
- Text alignment: identify reused text between two documents

From our perspective, we may be interested in the opposite of source retrieval:
- is the current document dissimilar from other ones written by the same author?

http://www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-web/plagiarism-detection.html
Deception detection

Deception detection, at least in one form, is the task of detecting whether a person is who they claim to be
- We are less interested in finding deceptive content by someone who we know who they are
- If we have user meta-data (e.g., age), deception detection could take the form of author profiling (e.g., does text match age?)

A similar task to plagiarism detection, but:
- There is an overt attempt to hide one's identity
- There is new content, i.e., you cannot compare content to other documents

One specific instance of this is sockpuppet detection: detecting whether a user account is a fake
- Can aggregate over multiple textual instances

Author profiling & attribution in social media

Author profiling: categorize authors based on a demographic property, such as:
- Sex/gender
- Age
- Political persuasion
- ...

Task is to take a document (or set of documents) and determine a category (e.g., "18–24 years old")
This is commonly of interest for work in social media, where demographic information can help with product assessment
http://www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-web/author-profiling.html

Dialect identification

Dialect identification: determine the (regional) dialect of a writer of a text
- A specific type of author profiling, often of interest to (socio)linguists

Discriminating between Similar Languages (DSL) task:
- determine which variety of a language (from among a set) a document belongs to
- i.e., which dialect is the writer using?

Unlike demographics such as age, note that there are fuzzy boundaries between dialects, and that various factors (demographics) interact
- i.e., language usage varies based on region, sex, age, ...

http://corporavm.uni-koeln.de/vardial/sharedtask.html

Native language identification

Native language identification: identify native language (L1) of writer based on second language (L2) writing
- Guessing a latent property based on systematic behavior in text
- As with dialect identification, we are getting into topics that are more specific to language usage
  - i.e., we are classifying a property of the writer's language, not some non-linguistic demographic

Many language learners have some similar patterns, e.g., (non)usage of articles
- But the entirety of their patterns tend to differ
- Note that data size is thus an issue (as with all these topics)

https://sites.google.com/site/nlisharedtask2013/home

Language proficiency identification

Language proficiency identification: determine the level of a language learner
- Given some level categorization (e.g., DLI, CEFR, course placement)
- Some researchers investigate criterial features, those features which best distinguish levels within a language

Language proficiency identification gets into more temporary & fluid properties of the writer
- proficiency today ≤ proficiency a year from now

Other topics?

And are there other topics to examine?
- I don’t know, but any (more or less permanent) demographic is in principle possible
  - where a demographic property is an inherent or cultural marker of the person
  - e.g., political persuasion, income, (language) disability, education, ...
  - nb: something like mood (or even political affiliation) is more fluid (a spectrum)

Fishing for some of these would probably lead to accidental correspondence, but we should be curious to find anything where:
- Input = natural language text(s)
- Output = classification of the writer, in terms of some demographic property
Where we’re going

For the next two-ish weeks, we’ll make sure we’re more or less on the same page:

- Machine learning & text classification techniques
- Natural Language Processing (NLP) tools
  - Our focus will be largely practical

After that, it’ll be less me & more you leading discussion

- By Monday, September 8, I want you to sign up to lead the discussion on a particular topic.
  - Lead ≠ Be an expert on
  - Lead = read papers ahead of time, help determine the interesting areas to explore, come to class with the interesting points identified, etc.
- I’ll give you a specific assignment on this next time

Expectations

This seminar is going to be:

1. Exploratory & interactive: think of this as collaborative learning in a lab-like environment
2. Demand-driven: I have only sketched a syllabus; the contents will be driven in part by your interests

But note that we can’t just pick one topic & ignore the others

- Notice how we have only scratched the surface on how these topics are similar and how they differ
- Issues such as inherentness, interaction of demographics, language-usage-specific, etc. are important to sort out
  - The degree to which we understand how to transfer techniques from one area to another depend on how well we understand this