Language and Computers (Ling 384)

Topic 3: Language Identification and Spam Detection

Adriane Boyd*
Department of Linguistics, OSU
Autumn 2005

The course was created by Markus Dickinson, Detmar Meurers and Chris Brew.

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 ö → 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 transiterated 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?

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, 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?

From language to spam identification

- The general idea of looking for recurring patterns of language carries over to identifying spam.
- spam = e-mail we don't want, usually only loosely directed to us, including unsolicited commercial e-mail
- Structure of discussion:
  - The issue and its social context
  - Language technology: rule and statistical methods
  - Devious spam
  - What you can do about spam

The issue

- Spam consumes
  - a significant fraction of total Internet bandwidth, which causes both a slowdown of other traffic, and possibly raises overall bandwidth cost.
  - a large amount of storage space on mail servers, sometimes actually making it temporarily impossible for "legitimate" messages to be received.
  - a significant portion of the time and effort of people who use email to communicate.
  - Spam can be the vehicle of "identity theft" campaigns, other types of fraud, and virus propagation.

(based on Spam: The Phenomenon by Colin Farhey, http://www.spiralsolutions.net/spam_topics/)
Two general types of language technology can be used for this:

- **Rule-based filters**
- **Statistical filters**

### Rule-based filters

This is basically **rule-based filtering** = filtering e-mail based on set rules. But rule-based spam filters can be more sophisticated:

- can weight patterns detected by the rules:
  - e.g., 3 points for viagra in the header, 2 for originating from a hotmail account, -2 points for a “.edu” address, ...
  - ⇒ When you pass some threshold of points, it’s marked as spam.
- can use information about systems it knows about:
  - e.g., This HTML message came from Outlook, but Outlook can’t send pure HTML messages

### Basic filtering

In setting up an e-mail account, you generally can set up the use of several folders and direct message accordingly.

- Send all mail with esp.com in the sender address to a separate sports folder. ⇒ Store messages you don’t need immediate access to.
- Delete all mail from viagra@spam.com ⇒ If you get mail from an address which never sends anything good (i.e., always spam), you never want to see it. You’ve effectively blacklisted it.
- Send all mail from my brother directly to my inbox. ⇒ Some messages you’ll always want to see right away. You whitelist these.

### Spam example

Spam detection software (here: spamassassin) has identified this incoming email as possible spam. It provides:

- **Content preview:**
  
  Email Marketing Email more than 2,500,000+ TARGETED prospects EVERYDAY! That’s over 75,000,000+ prospects per month (and growing!). Our Optin email safelists are 100% Optin and 100% legal to use. Your ad will reach only those prospects who have requested to be included in Optin safelists for people interested in new business opportunities, products and services. [...] 

- **Content analysis details:** (11.2 points, 5.0 required)

### Problems with Rule-based filters

Rule-based filters are quite intuitive and can be highly effective, but they also have drawbacks:

- Someone has to identify a pattern and specify a rule matching it (with high precision/recall).
- The more rules there are, the better it detects, but the slower it runs.
- Rule-based filters by nature are a step behind the spammers:
  - rules can only be developed once a pattern has been observed in spam, and
  - once a spammer knows a rule, they can try to bypass it.
Statistical filters

- Statistical filters have been proposed in place of or in addition to rule based ones.
- Instead of providing hand-written rules, one provides large sets of examples, one set with messages known to be spam, another with messages known to be ham.
- How it works:
  - Count up occurrences of words in previous e-mails:
    - How many times does X appear in something flagged as spam?
    - How many times does X appear in something which isn’t spam? (i.e., is ham)
  - From these counts, we calculate the spam probability of a word.

Detecting spam example

So, let’s say that you get an e-mail from me (or Anton, or a professor like Detmar) saying:

Hey, class, I just heard about a great opportunity in Nigeria to study and even make money.

I’ve also put the quiz on-line and asked one of the linguistics students to take it for a test drive so we can be pretty sure it works.

Anna (or Anton, or Detmar)

Example continued

- We extract words with high probabilities of being spam: opportunity, Nigeria, money, ...
- and words with low probabilities of being spam: linguist, Anton, Detmar [it’s hard to realistically fake an acquaintance’s name – though Anna’s common enough to be potentially problematic…]

We combine these probabilities, and it turns out that opportunity and money are indicators of spam, but Anton (or Detmar) and linguistics are very good indicators of non-spam.

Detecting spam

- We calculate this probability for every word.
- When a new e-mail comes in, we extract all the words and find their probabilities.
- We pick the 15 (or so) words which are the best and the worst indicators of spam (farthest from the middle) i.e., Pick the 15 words which give the strongest indication as to the true contents of the message.
- Combine these probabilities into a single probability
- If the probability is high enough (maybe 90% or more), call it spam.

Recalculating

Note that at some point, this non-spam e-mail will itself be used in recalculating probabilities for words.
- That is, the spam filter is continually learning what is spam and thus adapting to new spam techniques
- As with general document classification, this idea of machine learning is very important & widely-used.

Machine learning = computer learns how to behave based on previously-seen data.

Devious spam

- Spam filters try to distinguish spam from ham, using rules and patterns of word occurrences that it has learned about.
- Spammers want to disguise their messages so that they trigger none (or only few) of the rules and do not contain occurrences of words typical for spam.
- Emails are often encoded in HTML (hypertext markup language), so we need to talk about this encoding before we can take a closer look at various spammer tricks.

HTML

The Hypertext Markup Language (HTML) provides meta-information which tells a web browser or mail reader how a document is structured and how it should be displayed.

- HTML markup has beginning and end tags
  - <b>Example</b>: tells the browser to render the text Example in bold, i.e. as Example
  - An HTML tag can have attributes
    - For example, color is an attribute of the font tag.
    - <font color="blue">Language</font> makes Language appear blue
Tricks with spaces and characters

- Make words which are good indicators for spam look less like words:
  - Space out words to make them unrecognizable to word detectors e.g., M O R T G A G E
  - Other characters can be used instead to space things out e.g., P*R*E*E V I * A * G * R * A O!N L !$ N % E
- Spam detection software needs to keep up with spammers' tricks for encoding words.

Invisible Ink

- Spammers do things which can mess up your spam filter by secretly including words which make the e-mail sound legitimate, but which the e-mail user never sees.
  - Add some real random words before HTML:
    - suspensory obscure aristocratical meningorachidian unafeared brahmachari
  - Write white text on a white background:
    - e.g., milli <! xe64 >onaire
- Spam filters should include in their calculation exactly what the users sees.

Do you see what I see?

- One especially devious tactic involves taking English text and dividing it vertically:
  - Take the English text and instead of printing it out horizontally, print it vertically in a table
  - The result will look like English to the user, but will only be word fragments to the parser.
  - Again, filter needs to see what the human sees.

What to do?

- So, now that spammers are adding “good” words and hiding “bad” ones, what can we do?
  - Just throw our hands up and start looking into these great mortgage deals. :-(
  - Mix statistical filters (considers the good) and rule-based filters (still finds the bad).
  - Work to make sure that the filters see what the human sees.

Trick characters

- If you can alter characters, words won’t appear as the same words which are frequently found in spam:
  - Replace letters that look like numbers with numbers e.g., V1DE0 T4PE M0RTG4GE
  - Use accented characters in English e.g., Fántástic – earn ménésy through unqúõíëët judgedg
- Spam detection software needs to undo these mappings.

Split words with empty HTML tags

- Make it so that a single suspect word isn’t seen as a single word by the detector—but it is seen by the human as a single word.
  - e.g., milli<!xe64>onaire
  - Lesson: Filters are going to need to understand HTML very well.

Hiding the contents in other media

- Instead of encoding a message in a text, spammers:
  - send images
  - send http links to images
  - send <html>
  - send images which are good indicators for spam look less
  - send word fragments to the parser.

What you can do about spam

Negatives

- Don’t ever buy anything advertised through spam—if everyone observed this, spamming would not pay off and stop existing.
- Be careful about:
  - Asking to be taken off a list.
    - Clicking on “remove me,” or replying to spam mail will let them know your e-mail is valid.
  - Posting to a newsgroup which publicly archives their messages
  - Marking (or, more likely, not unmarking) that box when signing up for an account which says something like “I’d like to receive offers…”
  - Posting your e-mail on your website or in newsgroups.

Positives

- Things you can do:
  - Create accounts specifically used for newsgroups and such
  - Make your e-mail address on your website readable only to humans.
    - e.g., holbrook.1ATosuPERIOD—and don’t forget that “edu” at the end
  - Use a properly configured spam filter (e.g., the free spamassassin is very well configurable)