Text Analytics
Introduction (Part 1)

Maha Althobaiti, Udo Kruschwitz, Massimo Poesio

School of Computer Science and Electronic Engineering
University of Essex
udo@essex.ac.uk

23 September 2015
Brief Outline

- Part 1: Introduction to text analytics
  - Motivation
  - What is it?
  - State of the art
- Part 2: Some fundamentals
- Part 3: Glimpse into sample applications:
  - Web search
  - Question-answering
Example 1 - SIGNAL: Large-Scale News Analysis

Cam-puss kitty gets his own NUS card

A TOP cat unofficially adopted by Essex University has been given his own student card. Pebbles wanders into lectures and relaxes in the Teaching Centre on the Wivenhoe campus. The witty kitty been recognised as a key part of campus life to such an extent he has been registered as an honorary undergraduate with his own identity card. ... Original story

Cam-puss kitty gets his own NUS card

...
Example 1 - SIGNAL: Large-Scale News Analysis
Example II - Information Extraction

![EntityCube](entitycube.research.microsoft.com/view.aspx?id=9303)

**Related Organizations**

**Google**
- Issues running office xp on windows vista
- Google Maps Downloader
- Office 2009 cheap office 2010 home and business
- Buy [office](http://www.kickoffworld.net/blog/commodore-64-returns-on-nintendo-wii-virtual-console-ds-and-psp/)

**Nokia**
- How Nokia and Elop Will Fold Into Microsoft
- The return of prodigal Nokia CEO Stephen Elop to Microsoft's executive team as part of the acquisition of Nokia's handset and services unit is raising question...

**Novell**
- Steel Holdings, Boston Scientific, Coca-Cola, Dreamworks Animation, Estee Lauder, Home Depot, Microsoft, Novell, Procter & Gamble, Travelers, Viacom, Xerox and Yum! Brands

**Dell**
- WHQL Vista driver
- Microsoft Keyboard Driver
- Driver Dell 968 / 969w All-in-One Printer
- Driver Dell LifeCam Driver Dell Ageia PhysX Adapter Driver [Microsoft](http://Dell AIO G45 Interface...
Example III - Web Search

[Image of a computer screen showing a Google search page with a humorous pumpkin-themed design]
Example IV - Profile-Based Summarisation for Navigation

We focus on providing you with the skills and experience that you need in the graduate job market. Our courses are designed to enable you to enter a wide variety of careers and our graduates have gone on to work for several well-known employers.
Why is it Interesting?

- Web search engines
- Text classification
- Text summarization
- Filtering
- Human-computer dialogues
- Computer assisted language learning

... and more ...

- Generation (e.g. AI in games: check out Versu)
- Machine translation
- Speech processing
Why is it Difficult?

- Ambiguity
  
  *bank, book, ruler ...*
  
  I saw the man with the telescope on the hill.
  
  *Five monkeys ate three bananas.*
  
  *Time flies like an arrow.*
  
  *Lassen Sie uns noch einen Termin ausmachen.*

- Language is changing
  
  *I want to buy a mobile.*
  
  *Google it!*

- Ill-formed input
  
  *accomodation office*

- Multilinguality
  
  *Susanne Höll, Claudia Schiffer*
Two Paradigms

- Symbolic (deep) processing
  - hand coded rules
  - knowledge-rich

- Statistical (shallow) processing
  - build up statistical model of language
  - data-rich

... most powerful approaches nowadays are statistical!
NLE: Where we are

- Mature, everyday technology
  - E.g.: tokenization, normalization, regular expression search
- Solid technology that can still be improved
  - E.g.: lemmatization; spelling correctors; IR / Web search; speech synthesis
- Used in real applications, but needs improvement
  - E.g.: POS tagging; term extraction; summarization; speech recognition; text classification (e.g., for spam detection)
  - Spoken dialogue systems for simple information seeking (railways, phone)
- “Almost there” technology
  - E.g.: information extraction, generation systems, simple speech translation systems
- Pie in the sky
  - E.g.: full machine translation, more advanced dialogue
Text Analytics - Introduction (Part 2)

Linguistic Essentials;
Common Processing Pipelines
Levels of Linguistic Analysis

- **Word level:**
  - Parts of speech: DOG, EAT, RED
  - Sub-word level:
    - Phonetics, Phonology
    - Morphology

- **Phrase level (syntax):** THE RED DOG, CUTTING CORNERS, BY 3 O’CLOCK

- **Semantics:**
  - E.g., lexical semantics - COURSE / MODULE, DOG / ANIMAL

- **Discourse:**
  - E.g., anaphora - MOST STUDENTS SETTING OFF ON GAP YEAR TRIPS WILL NEED THEIR MONEY AND POSSESSIONS TO FIT SECURELY AND COMPACTLY INTO A RUCKSACK AND MONEYBELT
Words: Parts of speech (POS)

Words belong to different classes
- The {sad / sheep / *run / *always / *most} dog barked

Basic PARTS OF SPEECH:
- NOUN: dog, man, car, law
- ADJECTIVE: red, fat, brave
- VERBS: run, barked

Best known set of part of speech TAGS: Brown TAGS
- NN for nouns
- VB for verb base forms
- JJ for adjectives in positive form

Notice: many words belong to more than one class

Open and closed classes
Levels of linguistic processing: the basic pipeline of an NLP system (e.g., GATE)
Example: processing a query to a web search engine

List the estate agents in Stratford, London.

Text Analytics (Introduction II)
More advanced examples: Information Extraction Systems (e.g., LASIE)
In July 1995 CEG Corp. posted net of $102 million, or 34 cents a share.

Late last night the company announced a growth of 20%.
Preprocessing, I: tokenizing

In July 1995 CEG Corp. posted net of $102 million, or 34 cents a share.
Late last night the company announced a growth of 20%.
In July 1995 CEG Corp. posted net of $102 million, or 34 cents a share.

Late last night the company announced a growth of 20%.
Lexical Processing, I: POS tagging

CEG Corp. posted net of $102 million,
Lexical Processing, II: lemmatizing / stemming

CEG Corp. post net of $102 million,
An example of practical (partial) Parsing: Identifying numerical expressions

<W C='NNP'>CEG</W></W>
<W C='NN'>Corp.</W></W>
<W C='VBD'>post</W></W>
<W C='NN'>net</W></W>
<W C='IN'>of</W></W>
[NUMEX>
<W C='S'>$</W></W>
<W C='CD'>102</W></W>
<W C='NN'>million</W></W>
</NUMEX>
</W C='CM'>,
An example of practical semantic processing: identifying semantic type

CEG Corp. post net of $102 million,
An example of discourse processing: resolving anaphoric references

In July 1995 CEG Corp. posted net of $102 million, or 34 cents a share.

Late last night the company announced a growth of 20%.
Basic Pre-processing; Regular Expressions
The basic tasks in text processing

- **TOKENIZATION**: identify tokens in text
- **WORD COUNTING**: count words and their frequencies
- **SEARCHING FOR WORDS**
- **NORMALIZATION**:
  - UDO KRUSCHWITZ, udo kruschwitz, udo krushwitz → Udo Kruschwitz
  - Sep 23, 23rd of September, ..... → 23/09/2015
- **STEMMING**
Regular Expressions: a formalism for expressing search patterns

Because matching is a very common problem, over the years computer scientists have identified a set of patterns that

1. Are very common
2. Can be searched for efficiently

The language of REGULAR EXPRESSIONS has been developed to characterize these patterns.

Many programming languages (Perl, Java 1.4, TCL, Python… ) / web search tools / software systems (awk, sed, emacs) allow users to use regular expressions to specify what they are searching – these REs are then compiled into efficient code.

You do not need to write the code yourselves!
Types of regular expressions

Disjunction:
- `/centre|center/`
- `/accomodation|accommodation/`
- Also:
  - `/[Cc]entre/`
  - `/acco[m/mm]odation/`

Repetitions:
- `+`: Any number greater than 0
  - `/YES+/`
  - Matches YES!, YESS!, YESSS!
  - E.g., any binary number: `[01]`+
- `*`: 0 or more
  - `/ab*/`
  - Matches a, ab, abb, abbb
Applications of more complex REs

Web pages about Centres and Centers:

/Cc]entre\[Cc]enter/

Regular expression to validate phone numbers:

(+44)(0)20-12341234, 02012341234, +44 (0) 1234-1234
But not: (44+)020-12341234, 12341234(+020)
^\((?\+?[0-9]*\)?[0-9_\- \(\)]*\)$

Validating email addresses:

asmith@mactec.com, foo12@foo.edu, bob.smith@foo.tv
But not: asmith, @mactech.com, a@a
^([a-zA-Z0-9\-\.]+)@(([\[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}|([a-zA-Z0-9\-]+\.)+)[a-zA-Z]{2,4}|[0-9]{1,3})(\[\])*$
REs in action (linked to our own research)

Use of regular expressions as part of a processing pipeline to process query logs of a digital library (in this case: The European Library)

Why? We want to automatically learn what query modification suggestions to propose to (future) users of the library search engine, e.g.:
... example continued

- ...
  - 1889115 guest 71.249.xxx.xxx 8eb3bdv3odg9jncd71u0s2aff6 en ("mozart") search_url 0 2008-06-24 22:02:52
  - 1889118 guest 71.249.xxx.xxx 8eb3bdv3odg9jncd71u0s2aff6 en ("mozart") view_full 1 2008-06-24 22:03:03
  - 1889120 guest 71.249.xxx.xxx 8eb3bdv3odg9jncd71u0s2aff6 klavierkonzerte search_res_rec_all 0 2008-06-24 22:03:55
  - 1889121 guest 71.249.xxx.xxx 8eb3bdv3odg9jncd71u0s2aff6 ("klavierkonzerte") view_full 1 2008-06-24 22:04:10
  - ...

Text Analytics (Introduction II)
Here comes (part of the) the processing pipeline:

```
more $CleanedFile | gawk 'BEGIN {FS = " xxxx "} {if (($5=="en") && ($4 != "null") && ($6 !~ /^\([()]*\([()]*[^\[\]]*\)\[[()]*\]$/)) && ($7~/^search_/) && ($7 !~ /^search_adv/) {print $4 " xxxx " $1 " xxxx " $6 " xxxx " $13}}' | sort -k 1,1 -k 3,3n | gawk 'BEGIN {FS = " xxxx "} {if ((OLD_ID == $1 && (PRINT == "on"))) {print OLD_LINE; PRINT = "on"}; if ((OLD_ID != $1) && (PRINT == "on")) {print OLD_LINE; PRINT = "off"}; OLD_ID=$1; split($_, a, " xxxx "); OLD_QUERY = a[3]; OLD_LINE = $1}'} > $SessionQueries
```
This is what I get:

- 8eb3bdv3odg9jncd71u0s2aff6 xxxx 1889115 xxxx ("mozart") xxxx 2008-06-24 22:02:52

8eb3bdv3odg9jncd71u0s2aff6 xxxx 1889120 xxxx klavierkonzerte xxxx 2008-06-24 22:03:55

...
Basic Ideas of Statistical Approaches
Statistical NLE - Motivation

- So far: purely symbolic approaches
- But instead of hand-coding rules we could derive knowledge from text corpora automatically.
- Sample applications:
  - Spell checking
  - Handwriting recognition
- Example: How can Google predict the correct spelling of words?

"University off Essex" → Did you mean: "University of Essex"
Statistical Methods in NLE

Two characteristics of NL make it desirable to endow programs with the ability to LEARN from examples of past use:

- VARIETY (no programmer can really take into account all possibilities)
- AMBIGUITY (need to have ways of choosing between alternatives)

In a number of NLE applications, statistical methods are very common.

The simplest application: WORD PREDICTION
We are good at word prediction

Stocks plunged this morning, despite a cut in interest rates by the Federal Reserve, as Wall Street began ....
Statistics and word prediction

The basic idea underlying the statistical approach to word prediction is to use the probabilities of SEQUENCES OF WORDS to choose the most likely next word / correction of spelling error.

I.e., to compute

\[ P(w | W_1 \ldots W_{N-1}) \]

For all words \( w \), and predict as next word the one for which this (conditional) probability is highest.
Using corpora to estimate probabilities

But where do we get these probabilities? Idea: estimate them by RELATIVE FREQUENCY.

The simplest method: Maximum Likelihood Estimate (MLE). Count the number of words in a corpus, then count how many times a given sequence is encountered.

\[ P(W_1..W_n) = \frac{C(W_1..W_n)}{N} \]

‘Maximum’ because doesn’t waste any probability on events not in the corpus
Summary: Statistical Approaches

- Statistical methods typically need an annotated corpus (Maha ...)
- Methods can simply rely on counting (and a bit of smoothing)
- Same principle for any of your applications
- Very powerful!
- Use existing tools! (Massimo, Maha, ...)

Text Analytics (Introduction II)