From Text to Knowledge: Waving, not Drowning

Kalina Bontcheva
Department of Computer Science
University of Sheffield

6 September 2011
The Text Deluge

- 200 million Twitter users send 140 million tweets a day
- Facebook has:
  - 750 million active users, who
  - Spend over 700 billion minutes per month on the site
- Knowledge of the crowd
  - Buried in text on Wikipedia, Quora and other UGC
  - As well as in emails, publications, patents, etc.

Information Overload or Filter Failure (C. Shirky [http://bit.ly/oWJTNZ])
Text Analytics/NLP

- Text analytics is about knowledge discovery from large collections of unstructured text via application of Natural Language Processing (NLP) and analytical methods.
- Text analytics needs to address the unstructured nature of text and ambiguity of natural language.
- Market size in 2010: $835 million, with 25-40% growth potential (Seth Grimes, 2010)
T2K: GWAS Example

- Collaboration between Sheffield and WHO
- Derive knowledge about genes from papers
- Adjust the association priors used in Genome-Wide Association Studies (GWAS)
- A polymorphism ranked 124th using a standard odds ratios (OR) from patients and controls
- With TA - ranked 10th and considered highly relevant for a further study
- Half the data for the same results (saves 900 000 euros on a typical GWAS)
Towards Large-Scale T2K

- Cloud computing:
  - On-demand services to broaden research capabilities
  - Paradigm shift towards data-driven research

- NLP research still lagging behind:
  - 1 Million words typical corpus size
  - 500 Million tweets: 320h desktop, 30h server, 6h AWS
  - Several months to learn Hadoop, Amazon AWS and rewrite a part-of-speech tagging algorithm
  - NLP research methods need to change in order to deliver new large-scale TA methods
Towards Large-Scale T2K (2)

- NLP infrastructures abstract away common engineering tasks
- GATE (http://gate.ac.uk) – world-leading NLP infrastructure
  - Funded by EPSRC and EU research projects
  - Thousands of research users and over 100 SMEs and larger companies
  - Helps with document formats, evaluation, manual corpus annotation, and NLP module development
GateCloud.net

- A new cloud-based platform, based on GATE
- Research prototype supported by EPSRC/JISC
- Usual cloud benefits:
  - Pay-as-you-go, no upfront costs
  - No sysadmin work
  - Web-based management tools
- Not-so-usual cloud benefits
  - Based on open-source NLP software
  - Bring your own pipeline
  - No API required (all interaction is browser based)
GateCloud.net (2)

- **On demand servers** (always latest versions, backups)
  - Teamware: Collaborative text annotation tool
  - Mímir: Semantic Indexing and Search
- **On demand text processing**
  - Massive parallelism
  - On-line job definition tool
  - Many output formats, incl. XML & Mímir indexing
- **Processing tens of TB of text on Amazon EC2**
Semantic Search Example
(http://demos.gate.ac.uk/pin/)

- Politicians born in Sheffield, in Scottish news...

**People in the News**

**Looking For...**

Name: 

Fuzzy Name Matching

**Born In:** Sheffield

Famous As: Politician

**In Articles...**

Published Between 01/04/2011 and 30/04/2011

Classified As: Scotland

Ignore Boilerplate Text

**Results 1 to 2 of 2**

Scottish election: Respect Coalition Against Cuts profile
http://www.bbc.co.uk/news/uk-scotland-13048761 - Cached

... Bow - whose sitting MP Oona King had voted for the war ...

... success came when Galloway overturned Oona King's 10,000- ...
Challenge 1: T2K on the Cloud

**Large-scale T2K**
- Better algorithms
- Better models, derived from VL corpora

**Cloud-enabled NLP infrastructures**
- Lower the scaling up implementation overhead
- Make large scale TA affordable to small groups

**Shared repositories for large text collections**
- Repeatability of experiments
- Reducing costs (data is on cloud already)
- Anywhere availability
Challenge 2:
T2K from media streams

- Analysing media streams, in real-time, rather than batch, off-line processing
  - Text-based prediction modelling (markets and polls)
  - Need for incremental algorithms for real-time training and classification (online learning)
  - Mining non-stationary data: the statistical properties of relationships slowly change over time
- Semantic indexing and search over large text collections that update continuously
- Update NLP infrastructures for stream processing
Challenge 3: Analysing User-Generated Content

- Extracting knowledge from UGC is challenging
  - Conversational & Noisy
  - Personal
  - Contextual
  - Distributed

- Provenance and contradiction of opinions, capturing sentiment better (e.g. irony, sarcasm and slang)

- Understanding temporal expressions and modelling changes and dynamics over time
CGM Summarisation

200 new tweets
100 most recent messages
50 new photos

Automatic Summarisation

TA Symposium updates:
Forrester says TA not real time;
New classification method reported.

2 questions on LinkedIn about
your text analysis algorithms;
2 new jobs on NLP (details).

Mary agreed to meet 11am tomorrow here.

Personalisation

Interactivity

Learn from the user
Challenge 4: Annotation of Large Corpora

- Manual Annotation of Linguistic Corpora
  - $1M per million words of annotated text (OntoNotes)
  - 1 M word corpora too small; need 100 million or more
- Training of annotators
- Development of annotation guidelines
- Keeping annotators focused and motivated
Challenge 4 (2): Annotation of Large Corpora

- GATE Teamware: Collaborative, Web-based Annotation Environment

Process Monitoring: Annotation Status
Challenge 4 (3): Annotation of Large Corpora

- Crowd-sourcing (Mechanical Turk, CrowdFlower)
  - Similar quality, if many people annotate enough times
  - Only small scale corpora currently collected that way
  - Costs are lower, but only around 10 times so

- Further work needed:
  - Reusable NLP task breakdowns into HITs
  - Further experiments in crowd-sourcing very large corpora
Challenge 4 (4): Annotation of Large Corpora

- Games-with-a-Purpose (GWAP)
  - Side effect – corpus annotation
  - Used successfully for image annotation first
  - Some games for corpus annotation already created
  - Game design and incentives are specific to the kind of corpus required; no reuse; time-consuming
  - Very promising, further work needed
The Challenges In a Nutshell

1. T2K on the Cloud
   - Data-Driven, Affordable, Scalable Research
   - NLP Infrastructures and Repositories on the Cloud

2. Real-time Text Analytics from Streaming Media
   - Online learning, Non-stationarity, Semantic Search

3. Analysing Consumer-Generated Media
   - Robust, Adaptable, Cross-Media

4. Annotation of Very Large Corpora
   - Crowd-sourcing, GWAP
Thank you! Questions?

- For further details:
  - http://www.dcs.shef.ac.uk/~kalina/
  - http://gate.ac.uk
  - http://gatecloud.net
  - TrendMiner EU project (starting November 2011)
  - Arcomem project (http://www.arcomem.eu)