An Introduction to Text Mining
Research Papers

Petr Knoth
Phil Gooch

Mendeley
22 September 2015
What is text mining?

‘the process of deriving high-quality information from text’ (Wikipedia)

‘the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking ... of the extracted information ... to form new facts or new hypotheses to be explored further’ (Hearst, 2003)

‘a burgeoning new field that attempts to glean meaningful information from natural language text ... the process of analyzing text to extract information that is useful for particular purposes’ (Witten, 2004)
Why text mine research papers?

“Research papers are the most complete representation of human knowledge.”
Why is it so much talked about now?

The idea is not new, but up until recently access to large amounts of research papers was controlled by a handful of companies having bespoke arrangements with publishers.

The Open Access movement has recently largely contributed to decreasing the barriers to text-mining of research papers.

The availability of tools, developments in machine learning, and reduction in the costs of computing power and storage, has removed some of the technical and financial barriers.
What are the opportunities of text mining research papers?

- **Literature Based Discovery (LBD)**
  - Undiscovered Public Knowledge ([Swanson, 1986](#)).
  - Mining relationships for which there is “hidden” evidence in the research literature, yet they are not explicitly stated, such as *magnesium deficiency* and *migraine*, *fish oil* and *Raynaud's disease*.
  - Swanson's discoveries simulated by automated techniques ([Weeber et al., 2001](#)).
Other use cases

- Supporting exploratory access to research literature
  - staying up-to-date with research
  - analysing, comparing and contrasting research findings
- Summarisation of research findings
- Systematic literature review automation
  - ‘snowballing’
- Question answering and semantic search from papers
- Understanding the research impact of articles, individuals, institutions, countries, …
- Monitoring research trends
- Understanding how to direct research funding …
- Evidence of reuse and plagiarism detection …
Generic text mining workflow

1. Papers
   - Extraction and text cleaning
   - Tokenisation
   - Sentence splitting

2. Stemming, lemmatisation
   - Parsing
   - Part of Speech Tagging (POS)

3. Named Entity Recognition (NER)
   - Relation extraction

4. Knowledge base
   - Analysis and visualisation
   - Semantic search, reasoning, ...

(MENDELEY logo)
Example 1: Literature Based Discovery

- A range of techniques ([Smalheiser, 2012](#))
- A typical approach: ABC method (e.g. [Hristovski et. al. 2008](#)):
  - A affects/binds/regulated/interacts with B
  - B affects/binds/regulated/interacts with C
  - A and C are not explicitly linked in any article

=> There might be an undiscovered relationship between A and C

Various reasons why A and C might not be connected, e.g. B is a rare term.
Example 1: Literature Based Discovery

Papers on small-cell lung carcinoma

Term and relation extraction

List of terms

Search engine

Papers resulting from term queries

Recognise entities, such as symptoms, conditions, therapeutic agents (drugs)

List of terms

Term and relation extraction

Lists B and C with semantic annotations

Knowledge base

Ranked list of hypotheses connecting A and C

Medical researcher
Example 1: Visualising the A-C hypotheses

Reproduced with permission from FastVac and Public Health England.
Example 2: modeling research arguments using citation contexts


- **Section: Background** -- ... ities (e.g., proteins) reported in text. Approaches to the relation extraction task can be categorized into two major classes: (1) machine learning-based approaches and (2) pattern-based approaches. Machine learning-based approaches are data-driven and can derive models from a set of annotated data \([1-7]\). The use of machine learning methods can be quite effective, but the performance of resulting systems depends on the quality and the amount of annotated data. For example, large annotated corpora become available for the protein–protein interaction relation extraction @ 2014 Bao et al.
Example 2: modeling research arguments using citation contexts

- Label the citation network according to the context of how they are cited
- Pioneering work by Simone Teufel in this area

Green: Contrastive or comparative statements
Pink: Statements about the aim of the paper

Example 3: exploratory search over research papers

How can I get started?

- Get the data
- Design/build your workflow
- Select a framework, tools, services to be used in implementing the workflow
- Understand how to evaluate the performance of each component
Full-text Open Access article sources

- **Subject repositories:**
  - arXiv bulk data
  - PubMed OA Subset

- **Institutional repositories:**
  - ~3k across the world, see Directory of Open Access Repositories

- **Open Access journals:**
  - >10k OA journals, see Directory of Open Access Journals (DOAJ)

- **Open Access subsets from publishers:**
  - Elsevier OA STM Corpus
  - SpringerOpen

- **Aggregators:**
  - CORE (API, Data dumps)
Other full-text sources

Typically for non-commercial research only.

- Publisher full-text APIs:
  - Elsevier’s Text and Data Mining API:
  - (subscribed content on ScienceDirect)
- Domain specific corpuses:
  - JSTOR Data For Research
- Full-text content from multiple publishers:
  - CrossRef Text and Data Mining API (full-texts)
  - Mendeley API (abstracts)
- Other useful scholarly resources:
  - Microsoft Academic Graph (MAG)
- Clinical records
  - i2b2 NLP Challenge Data Sets
Building a text mining application

Papers → Extraction and text cleaning → Tokenisation → Sentence splitting

Stemming, lemmatisation → Parsing → Part of Speech Tagging (POS)

Named Entity Recognition (NER) → Relation extraction

Knowledge base → Analysis and visualisation

Semantic search, reasoning, ...
Example: Part of speech tagging


Example: Parsing

[Acute lymphoblastic leukemia]_{NP} ([ALL]_{NP}) [leads to]_{VP} [an]_{DT} [accumulation]_{NN} [of]_{IN} [immature lymphoid cells]_{NP} [into]_{IN} [the]_{DT} [bone marrow]_{NP}, [blood]_{NP} [and]_{CC} [other organs]_{NP}.

[A]_{DT} [young patient]_{NP} [was treated unconventionally for]_{VP} [Philadelphia positive ALL]_{NP} [and]_{CC} [Mucormycosis]_{NP} [with]_{IN} [Amphotericin B]_{NP}.

[The]_{DT} [patient]_{NP} [achieved]_{VP} [a]_{DT} [disease-free survival]_{NP} [of]_{IN} [12 months]_{NP} [with]_{IN} [good]_{JJ} [quality of life]_{PP}.
Example: Named entity recognition

[Acute lymphoblastic leukemia]_Disease (ALL)_Disease [RESULT]_VP [accumulation]

[of]_IN [immature lymphoid cells]_AnatomicalSite [into]_IN [the]_DT [bone marrow]

[young patient]_Person [TREAT]_VP [Philadelphia positive ALL]_Disease [and]_CC

[Mucormycosis]_Disease [with]_IN [Amphotericin B]_Treatment

[The]_DT [patient]_Person [RESULT]_VP [disease-free survival]_Outcome [of]_IN [12 months]

[Duration]_IN [good quality of life]_Outcome
Example: Named entity recognition

Acute lymphoblastic leukemia (ALL) leads to an accumulation of immature lymphoid cells into the bone marrow, blood and other organs. A young patient was treated unconventionally for Philadelphia positive ALL and Mucormycosis with Amphotericin B.

The patient achieved an overall and disease free survival of 12 months.

- Here, ALL is automatically expanded to its full form: *acute lymphoblastic leukemia*
- ALL is automatically labelled as a DiseaseOrSyndrome. as the initial mention was also labelled as DiseaseOrSyndrome
Example: Named entity recognition

- *Philadelphia* on its own can be labelled as a **Location**
- In this context it is part of a longer noun phrase ‘Philadelphia positive ALL’, of which **ALL** is the Disease **acute lymphoblastic leukemia**
- ‘*Philadelphia positive*’ is also a term in a domain-specific dictionary
- Which label is chosen can be determined by a rule, either handwritten or learnt by a machine learning algorithm
Frameworks

- General Architecture for Text Engineering (GATE)
- Apache UIMA
- Natural Language Toolkit (NLTK)
- OpenNLP
Tools and Services

- **Open Calais**: general purpose tagging of people, places, companies, facts, events and relationships
- **AlchemyAPI**: similar to OpenCalais
- **National Centre for Text Mining**: terms, entities, semantic search
  - **Cafetiere**: text mine your own documents online
  - **EventMine**: relations between terms: protein binding, synthesis inhibition
- **ContentMine**: Fact extraction
- **ReVerb**: open-domain relation extraction
- **XIP Parser**: linguistic analysis
- **Linguamatics**
- Metadata and citation extraction:
  - **ParsCit**
  - **Grobid**
  - **Cermine**
Open source demos

- **Mimir**
  - analysing, comparing and contrasting research findings

- **EEXCESS**
  - Recommend related resources
  - Narrative paths between resources
Existing challenges

- Harmonising metadata formats across publishers, repositories, etc.
- Agreeing on standards/ontologies/formats used to share the outputs of research publications text-mining tools and all their components.
- Integration of text-mining tools with content providers’ systems
- Building and maintaining text-mining web-services for research (building blocks)
- Promoting and adopting end-user tools utilising text-mining in researchers’ daily workflows
- Building gold standards for various text-mining tasks and sharing them across researchers (issue of credit)
Events and initiatives

- **Events**
  - International Workshop on Mining Scientific Publications (WOSP)
  - Joint Conference on Digital Libraries (JCDL)
  - ACL, COLING, IJC-NLP, CoNLL, LREC, etc.

- **Projects**
  - OpenMinTeD (EC funded)
  - EEXCESS
  - OpenAIRE
Thanks for listening ...

data-science@mendeley.com