An Introduction to Text Mining Research Papers

Petr Knoth
Phil Gooch

Mendeley
22 September 2015

MENDELEY

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*)

MENDELEY

Why text mine research papers?

“Research papers are the most complete representation of human knowledge.“

MENDELEY

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.

MENDELEY

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).

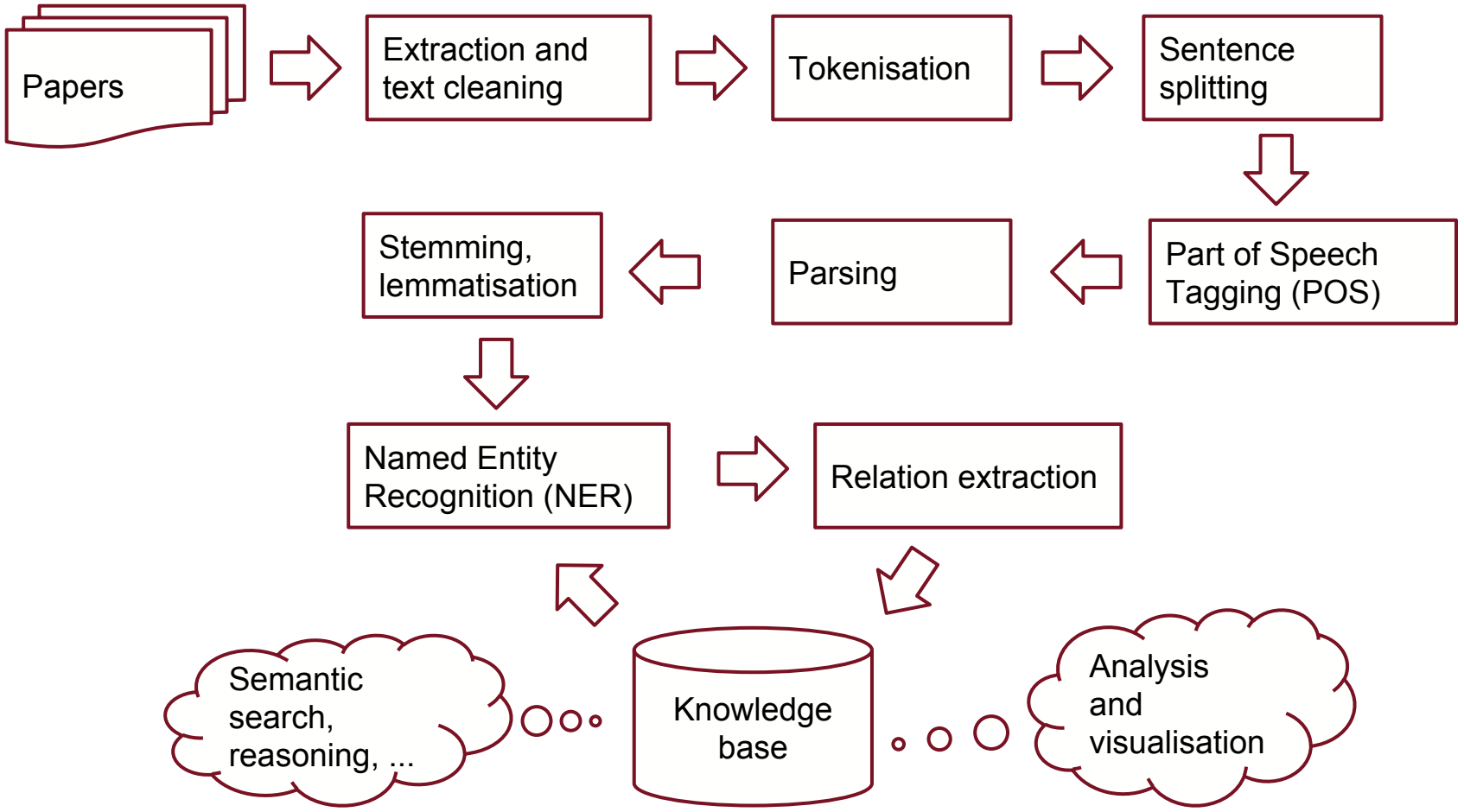
MENDELEY

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 …

MENDELEY

Generic text mining workflow



MENDELEY

Example 1: Literature Based Discovery

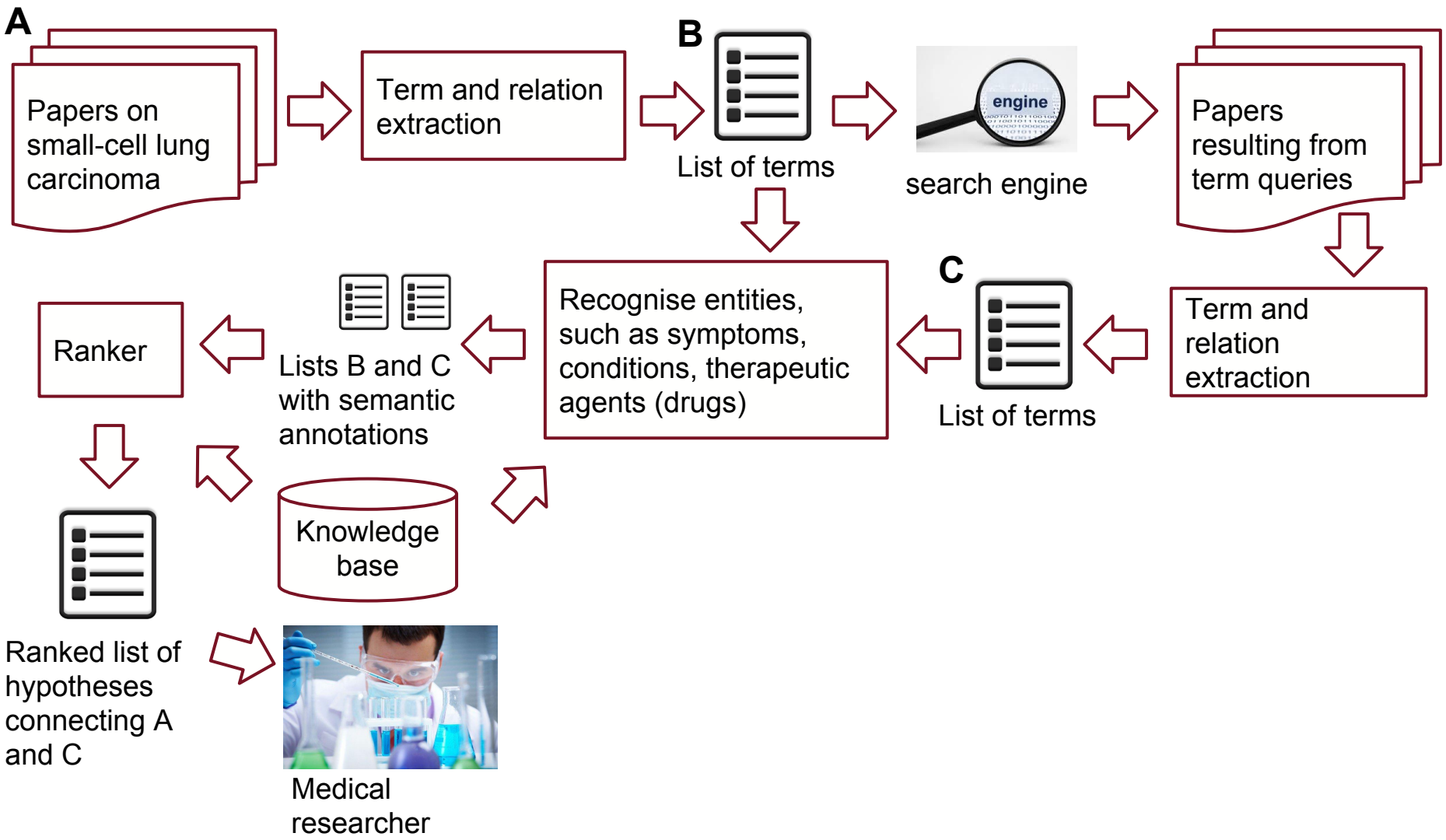
- A range of techniques (Smalheiser, 2012)
- A typical approach: ABC method (e.g. Hristovski et. al. 2008):
  - **A** affects/binds/regulates/interacts with **B**
  - **B** affects/binds/regulates/interacts with **C**
  - **A** and **C** are not explicitly linked in any article

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

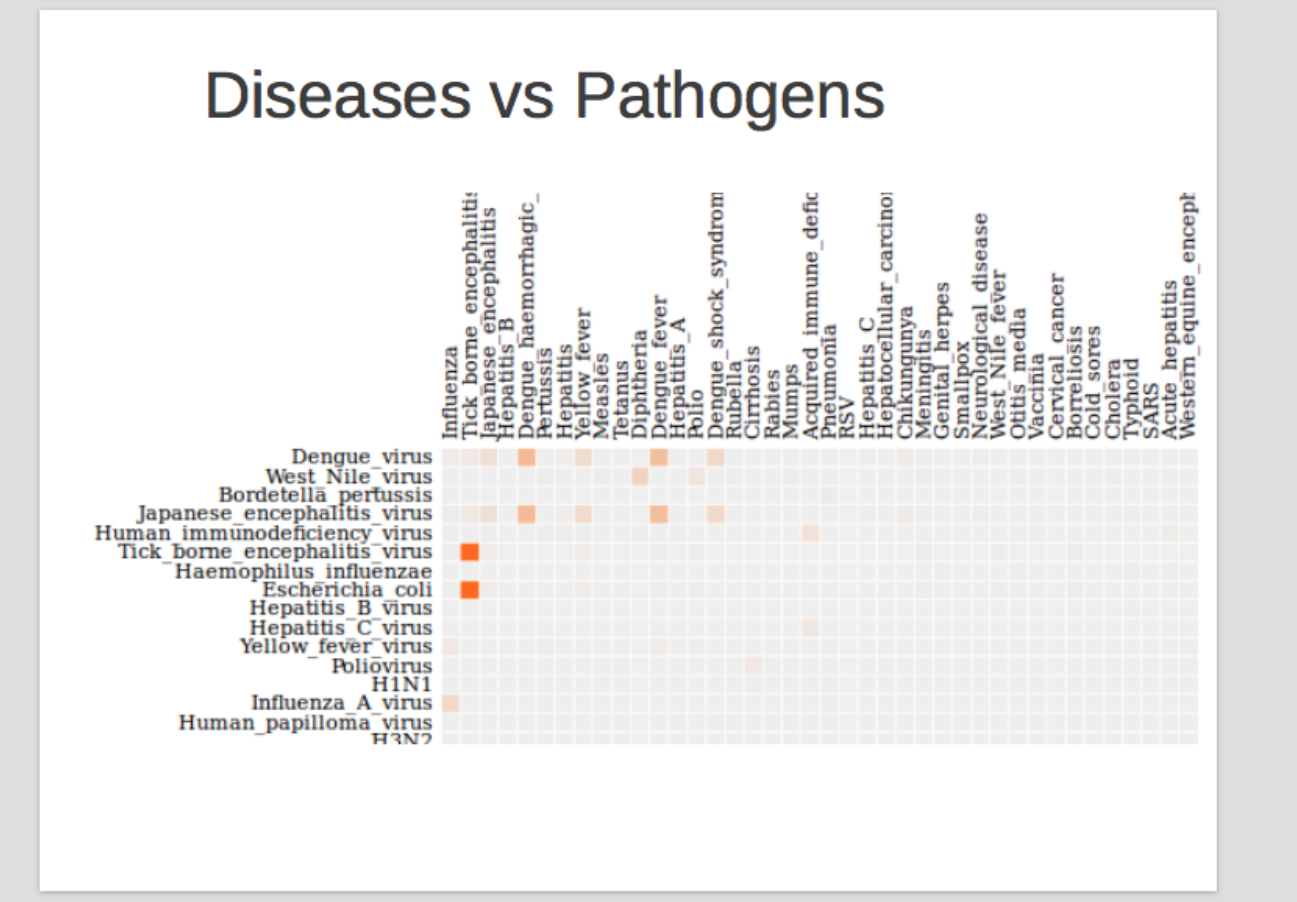
Various reasons why **A** and **C** might not be connected, e.g. B is a rare term.

MENDELEY

Example 1: Literature Based Discovery



Example 1: Visualising the A-C hypotheses

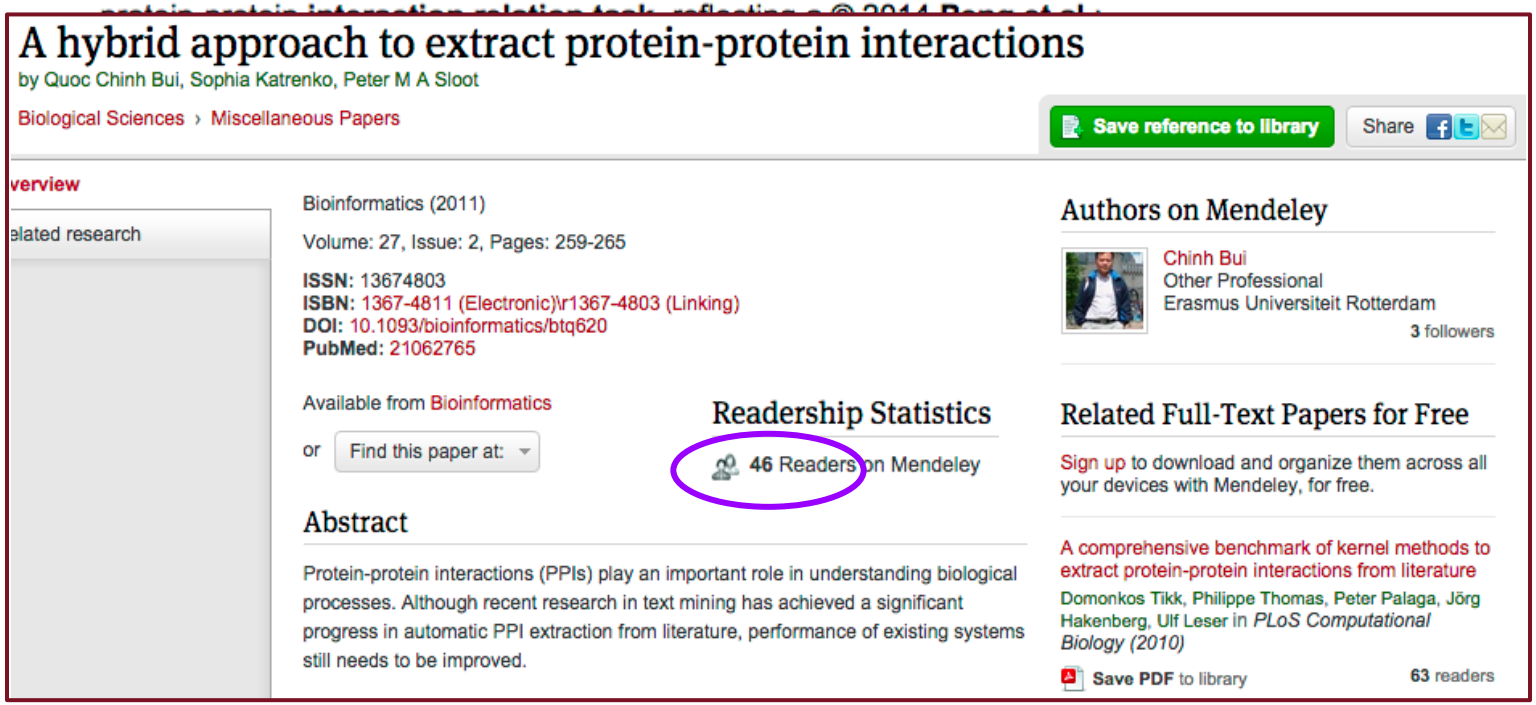


Reproduced with permission from FastVac and Public Health England.

Example 2: modeling research arguments using citation contexts

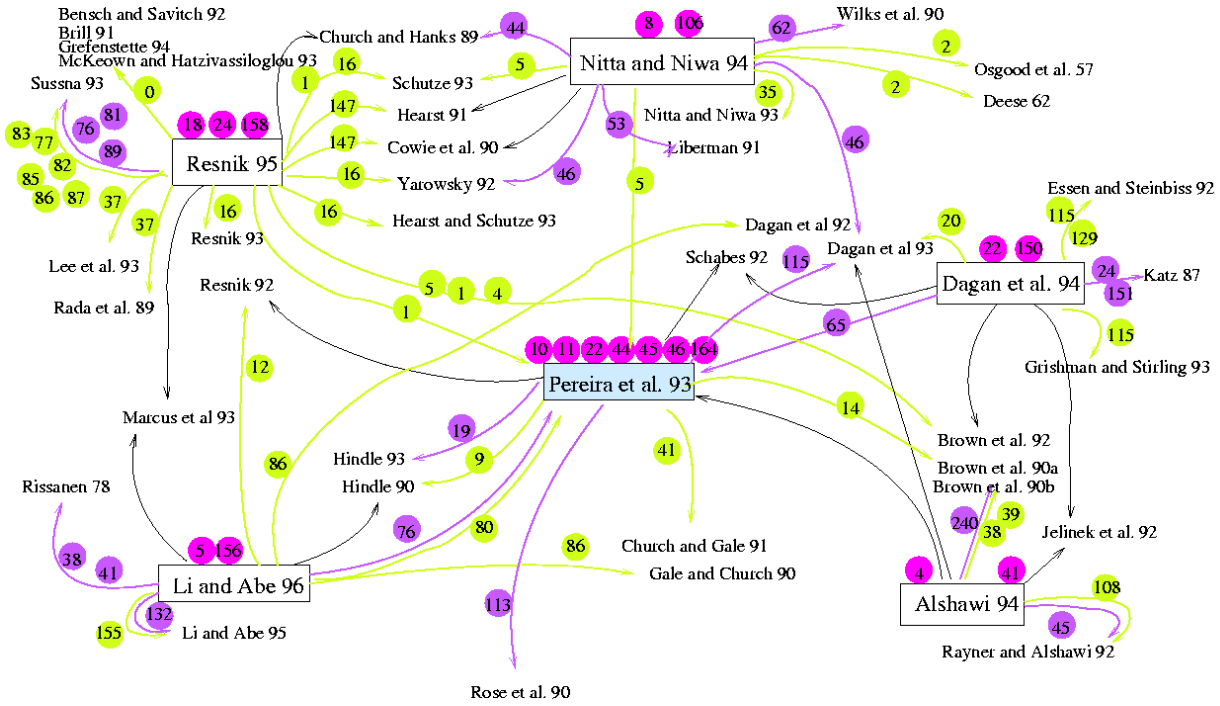
4. Bui QC, Katrenko S, Sloot PM: A hybrid approach to extract protein–protein interactions. Bioinformatics 2011, 27(2): 259–265. [Mendeley]

- **Section:** *Background* -- ... ities (e.g., proteins) reported in text. Approaches to the **relation extraction task** can be categorized into two major classes: (1) machine learningbased 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



Example 2: modeling research arguments using citation contexts

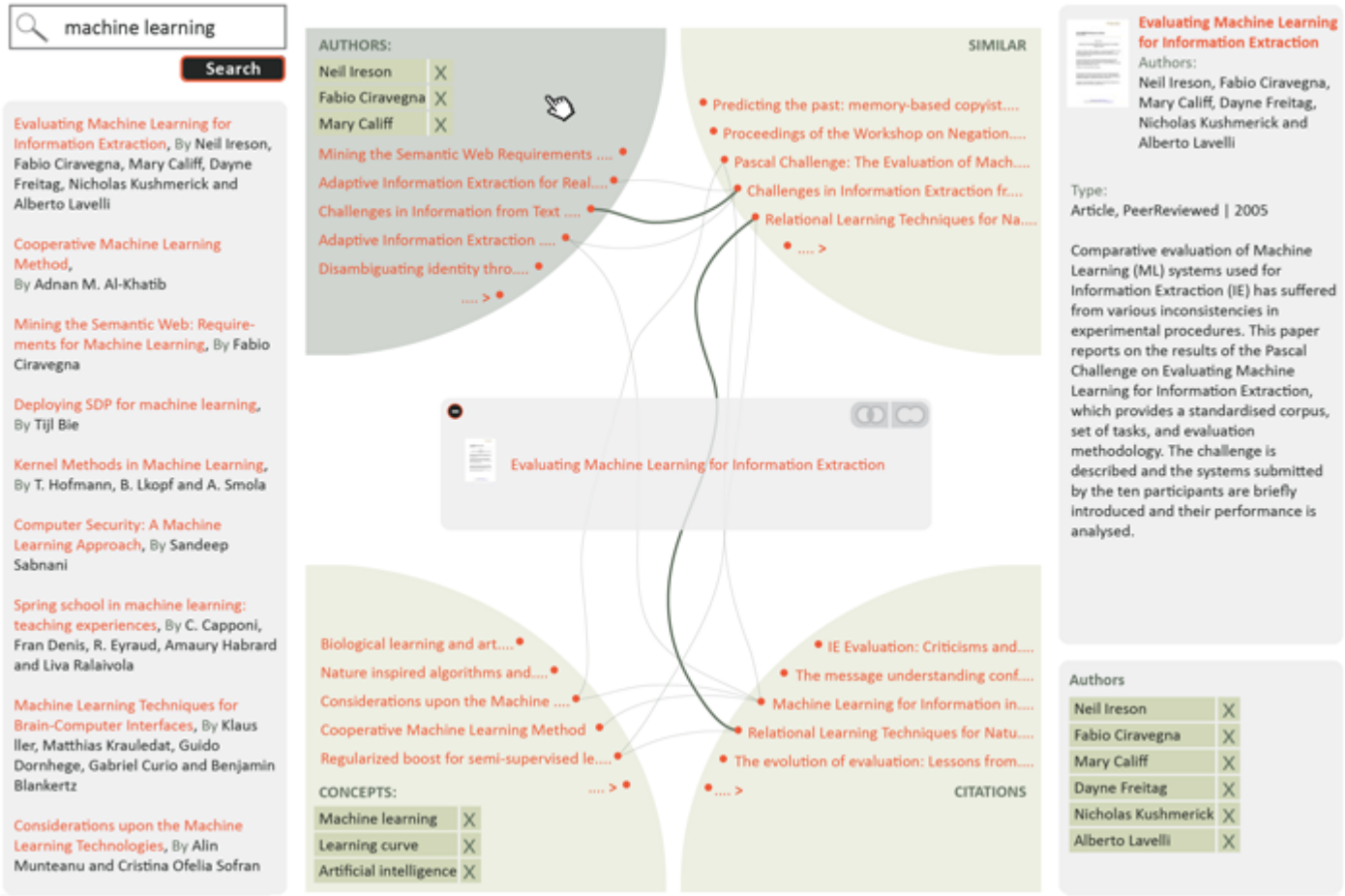
- Label the citation network according to the context of how they are cited
- Pioneering work by Simone Teufel in this area



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

Source: Simone Teufel (1999) Argumentative Zoning: Information Extraction from Scientific Text, Phd Thesis, University of Edinburgh.

Example 3: exploratory search over research papers



Source: Herrmannova & Knoth (2012) Visual Search for Supporting Content Exploration in Large Document Collections, DLib 18(8).

MENDELEY

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

MENDELEY

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)

MENDELEY

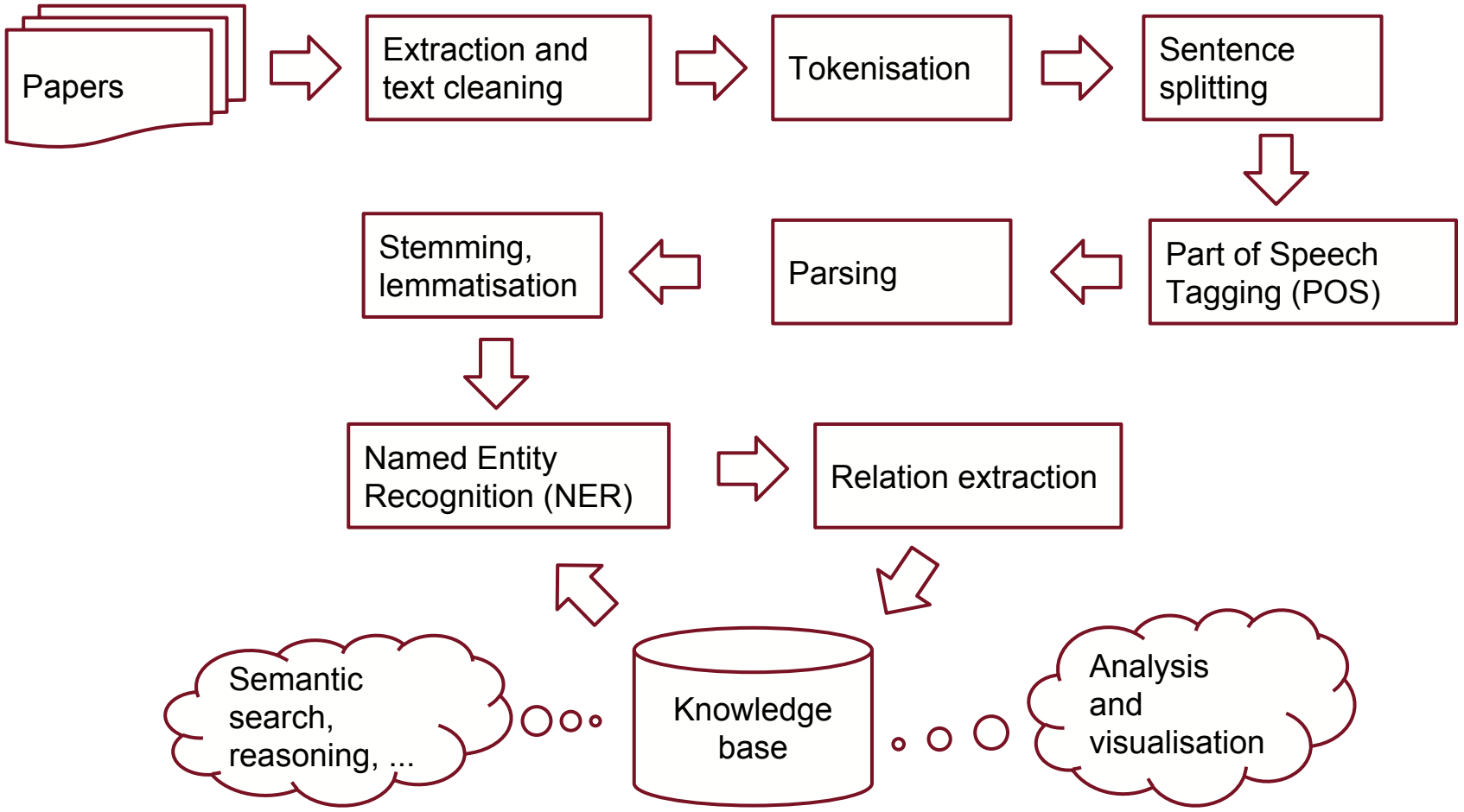
Other full-text sources

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

MENDELEY

Building a text mining application



MENDELEY

Example: Part of speech tagging

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

[A]$_{DT}$ [young]$_{JJ}$ [patient]$_{NN}$ [was]$_{VB}$ [treated]$_{VB}$ [unconventionally]$_{RB}$ [for]$_{IN}$ [Philadelphia]$_{NN}$ [positive]$_{NN}$ [ALL]$_{NN}$ [and]$_{CC}$ [Mucormycosis]$_{NN}$ [with]$_{IN}$ [Amphotericin]$_{NN}$ [B]$_{NN}$.

[The]$_{DT}$ [patient]$_{NN}$ [achieved]$_{VB}$ [a]$_{DT}$ [disease-free]$_{JJ}$ [survival]$_{NN}$ [of]$_{IN}$ [12]$_{CD}$ [months]$_{NN}$ [with]$_{IN}$ [good]$_{JJ}$ [quality]$_{NN}$ [of]$_{IN}$ [life]$_{NN}$.

MENDELEY

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]$_{PP}$.

MENDELEY

Example: Named entity recognition

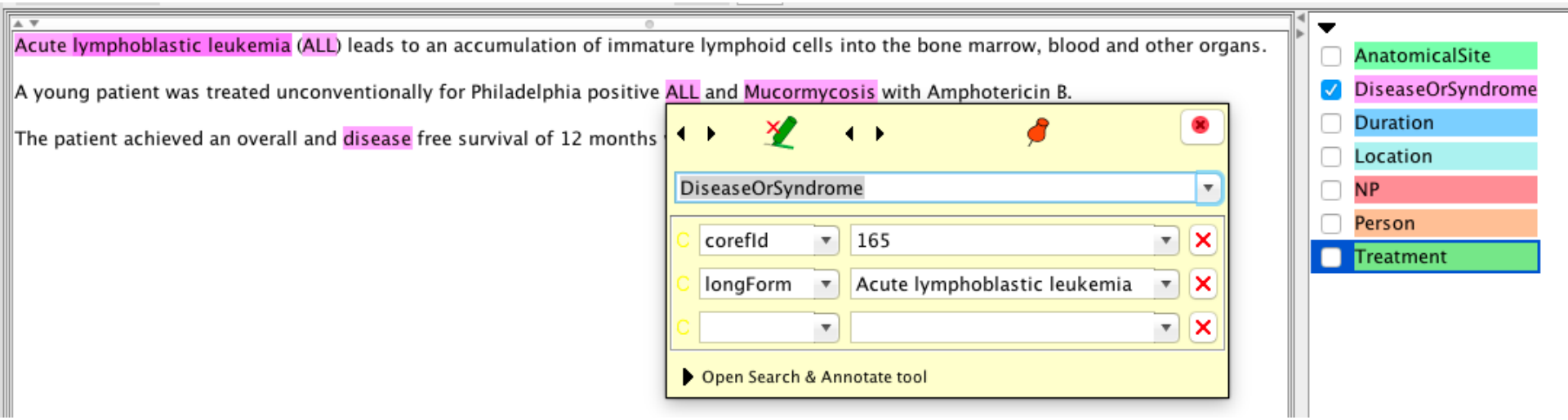
[**Acute lymphoblastic leukemia**]$_{\text{Disease}}$ ([**ALL**]$_{\text{Disease}}$) [**RESULT**]$_{\text{VP}}$ [accumulation]$_{\text{NN}}$ [of]$_{\text{IN}}$ [immature lymphoid cells]$_{\text{AnatomicalSite}}$ [into]$_{\text{IN}}$ [the]$_{\text{DT}}$ [bone marrow]$_{\text{AnatomicalSite}}$, [blood]$_{\text{AnatomicalSite}}$ [and]$_{\text{CC}}$ [other organs]$_{\text{AnatomicalSite}}$.

[A]$_{\text{DT}}$ [*young patient*]$_{\text{Person}}$ [**TREAT**]$_{\text{VP}}$ [**Philadelphia positive ALL**]$_{\text{Disease}}$ [and]$_{\text{CC}}$ [**Mucormycosis**]$_{\text{Disease}}$ [with]$_{\text{IN}}$ [**Amphotericin B**]$_{\text{Treatment}}$.

[The]$_{\text{DT}}$ [*patient*]$_{\text{Person}}$ [**RESULT**]$_{\text{VP}}$ [disease-free survival]$_{\text{Outcome}}$ [of]$_{\text{IN}}$ [12 months]$_{\text{Duration}}$ [with]$_{\text{IN}}$ [good quality of life]$_{\text{Outcome}}$.

MENDELEY

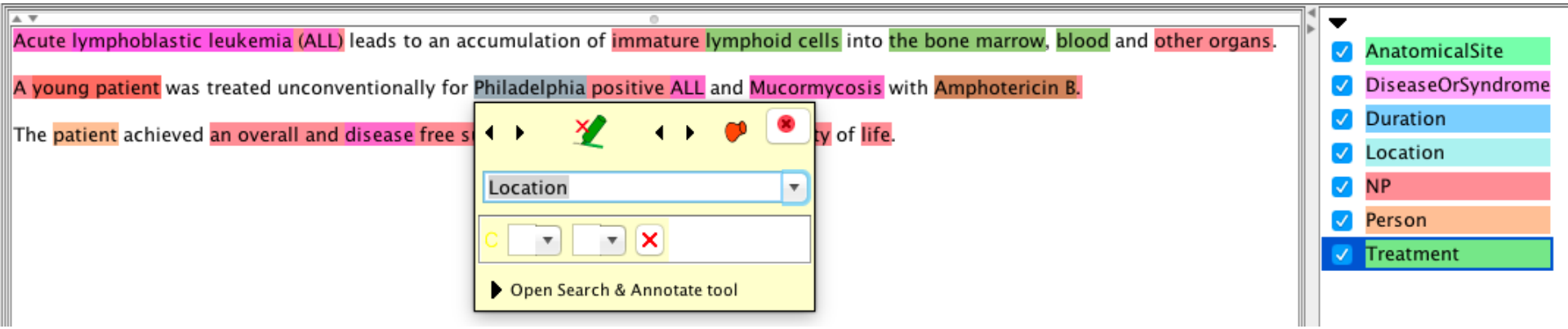
Example: Named entity recognition



- 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

MENDELEY

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

MENDELEY

Frameworks

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

MENDELEY

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

MENDELEY

Open source demos

- Mimir
  - analysing, comparing and contrasting research findings
- EEXCESS
  - Recommend related resources
  - Narrative paths between resources

MENDELEY

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

MENDELEY

Thanks for listening ...

datascience@mendeley.com