ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature

Callum Court
Molecular Engineering, University of Cambridge
Supervisor: Dr Jacqueline Cole

Overview

1. Introduction
2. Previous work
3. Challenges
4. Overview of the ChemDataExtractor toolkit
5. Applications

Introduction

- Approximately 20,000 new compounds and properties published in 10,000 biomedical chemistry journals in 2013 alone[1]
- Ideally we would compile all available scientific data into a database of material properties
- Too much data to extract manually

Introduction

- Scientific results are typically presented in papers, patents, these etc
- Containing unstructured and semi-structured data in the form of text, tables, captions and figures not readily interpretable by machines
- Modern Machine Learning and Natural Language Processing (NLP) techniques provide us with the means for automated information extraction

Previous work

- Large scale data-mining for materials discovery:
  - The Materials Genome Initiative[2]
  - The Harvard Clean Energy Project[3]
  - The Materials Project[4]
- Text mining tools for the Chemistry domain:
  - ChemicalTagger[5]
  - ChemEx Project[6]

Previous work

- Previous methods tend to focus on on predicting chemical properties confined to a particular field of research (photovoltaics, batteries etc.)
- All would be well complemented by a generic method for generating databases of materials properties in a domain-independent way

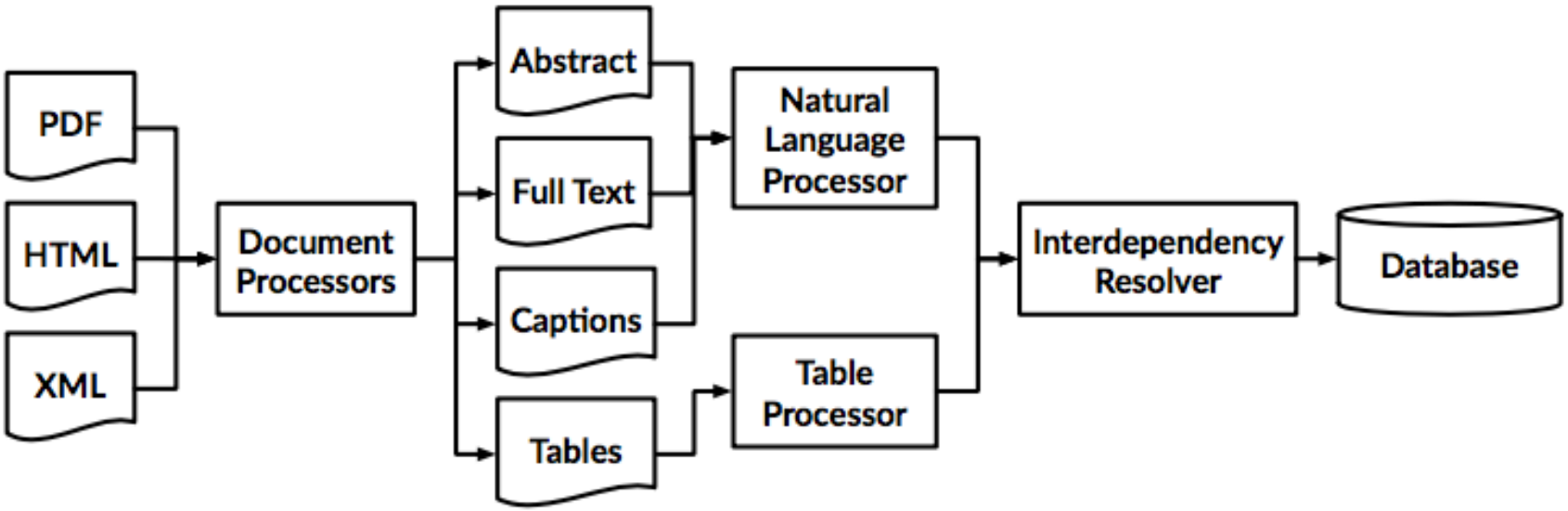
Challenges

- Although the scientific literature is relatively formulaic and structured, text-mining the scientific literature is very difficult
  - Each sub-domain of science has its own specific terminology and abbreviations
  - These conventions can vary between papers (and perhaps between sections)
  - Each sentence/paragraph cannot be processed individually as information is spread out through multiple sections

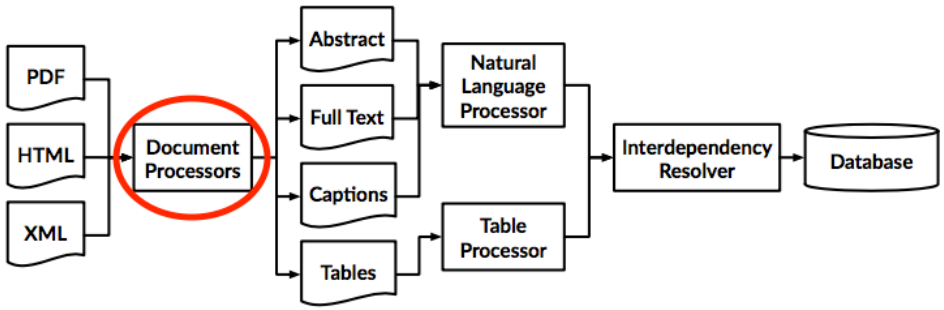
ChemDataExtractor (CDE)

- A comprehensive toolkit for the automated extraction of chemical information from scientific documents.
- Full extraction of melting points, glass transitions, UV-Vis absorption spectra and more
- Full source code and documentation available under MIT license at `www.chemdataextractor.org`

ChemDataExtractor (CDE)

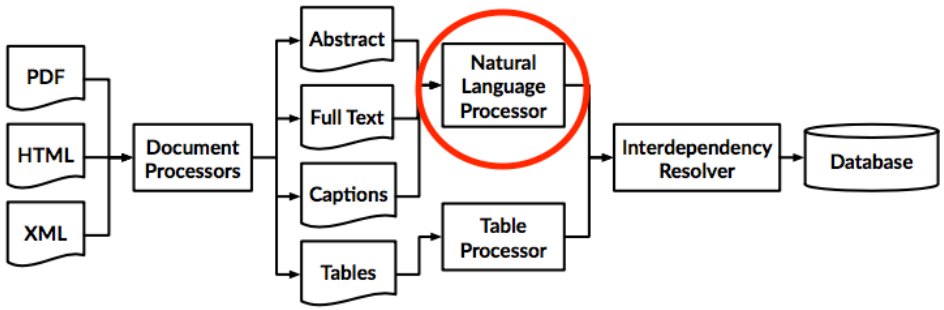


Document processing



- This stage converts differing file types into a single consistent structure consisting of abstracts, paragraphs, figures, captions and tables
- Enables all subsequent stages to perform in the same way regardless of initial document type

Natural language processing



- The key stage of the CDE pipeline where relationships and information are extracted from the text of the document
  1. Tokenization
  2. Part-of-speech tagging
  3. Entity recognition
  4. Phrase parsing
  5. Information extraction

Natural language processing

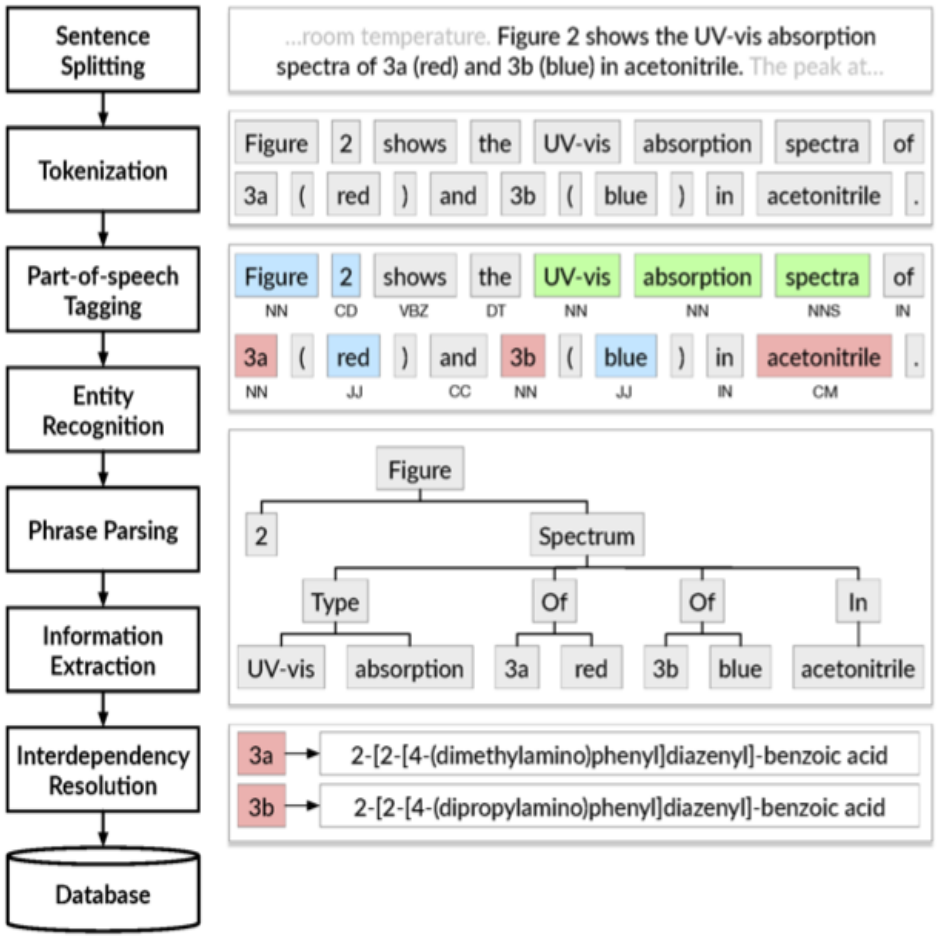
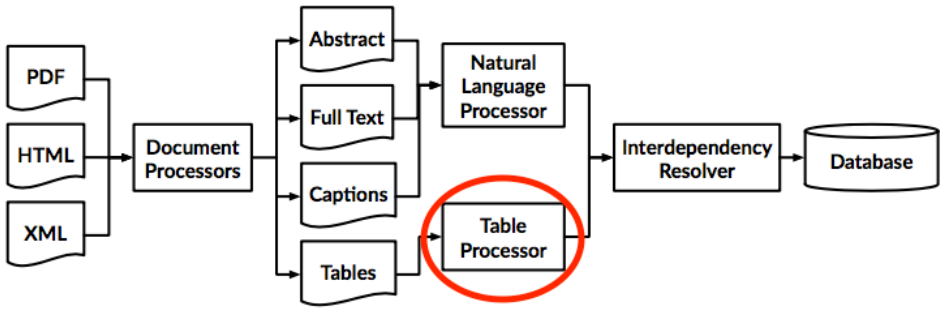
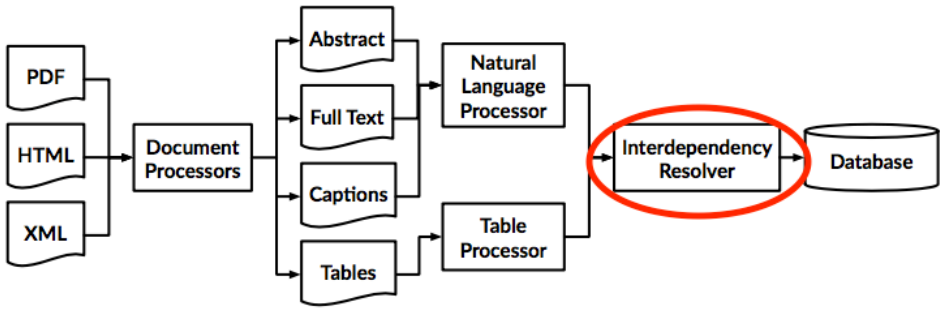


Table parsing



- Tables are an ideal source for retrieving structured data
- This stage treats tables as highly condensed forms of text
- Specialised rules are used to parse table headers and columns in the same way as normal text

Interdependency resolution



- Finally, all information from the natural language processor and table processor can be brought together
- This stage resolves the interdependencies between different sections and compiles all information into a set of structured records
- These records can be easily compiled into a database

Performance

	Precision	Recall	F-Score
Chemical identifier records	94.1%	92.7%	93.4%
Spectrum records	88.3%	85.4%	86.8%
Chemical property records	93.5%	89.6%	91.5%

- Evaluation performed on a set of 50 chemistry articles sourced from the Royal Society of Chemistry, American Chemical Society and Elsevier
- Precision: The fraction of retrieved records that are correct
- Recall: The fraction of correct records that are retrieved
- F-score: The harmonic mean of Precision and Recall

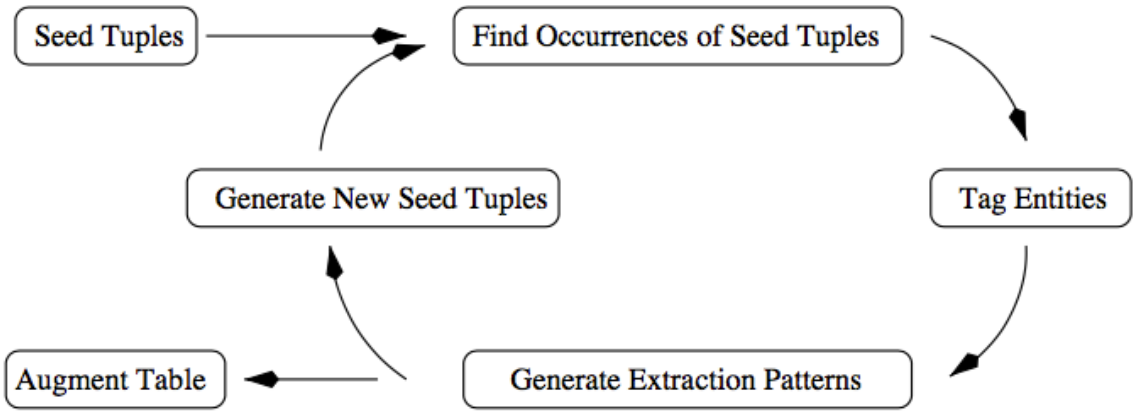
Applications

- Autogenerated databases of material properties can have great utility in materials science:
  1. Materials or drug discovery
  2. Property prediction
  3. Compound identification
  4. Research design

Current work

- Currently work is being undertaken to enhance the capability of CDE to extract properties associated with the physics corpora
- In particular, the extraction of magnetic properties with the aim of creating large auto-generated databases of magnetic properties

The Snowball algorithm



- The rule-based approach to phrase parsing is highly inefficient
- The Snowball algorithm[7] is a semi-supervised machine learning approach to probabilistic phrase parsing
- Initial results demonstrate a large increase in precision and F-score for CDE when a Snowball step is included into the pipeline

Summary

- ChemDataExtractor provides a complete pipeline for automatically extracting chemical data from the scientific literature in a domain independent way
- The overall system presents a high F-score of over 90% when applied to the chemistry literature
- Further enhancements to the system may be able to push this score even higher and make the system more suited for use in the physics domain
- This has great potential for use in materials science research

References

[1] Rmy D Hoffmann, Arnaud Gohier, and Pavel Pospisil. *Data mining in drug discovery*. Wiley-VCH, 2013. Chap. 5.

[2] National Science and Technology Council (US). *Materials genome initiative for global competitiveness*. Executive Office of the President, National Science and Technology Council, 2011.

[3] Roberto Olivares-Amaya et al. "Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics". *Energy & Environmental Science* 4.12 (2011).

[4] Anubhav Jain et al. "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation". *APL Materials* 1.1 (2013).

[5] Lezan Hawizy et al. "ChemicalTagger: A tool for semantic text-mining in chemistry". *Journal of cheminformatics* 3.1 (2011).

[6] Atima Tharatipyakul et al. "ChemEx: information extraction system for chemical data curation". *BMC bioinformatics* 13.17 (2012).

[7] Eugene Agichtein and Luis Gravano. "Snowball: Extracting relations from large plain-text collections". *Proceedings of the fifth ACM conference on Digital libraries*. ACM. 2000.