Natural Language Processing for Intelligent Access to Scientific Information

Francesco Ronzano and Horacio Saggion

Natural Language Processing Group (TALN)
Universitat Pompeu Fabra, Barcelona, Spain

Tutorial @ COLING 2016
11th December 2016
TALN @ UPF

Natural Language Processing Group

• Francesco Ronzano, Senior Post-doc
• Horacio Saggion, Research Professor
What it’s all about

Scientific literature is growing at an unprecedented rate

Automated approaches to extract, enrich, aggregate and summarize information from scientific publications are essential to enable any careful and comprehensive assessment of scientific literature

Natural Language Processing and Text Mining play a fundamental rule since they are key technologies to analyze scientific publications

This tutorial provides an overview of the core content analysis challenges and opportunities of Scientific Literature Mining showing how we can characterize and take advantage of implicit and explicit traits of scientific publications to better organize and provide access to scientific literature

http://taln.upf.edu/pages/coling2016tutorial/
Outline

• SCIENTIFIC INFORMATION OVERLOAD
  How much scientific literature is there out there? How can we search for and access to scientific information?

• DOCUMENT STRUCTURE ANALYSIS
  How can we extract textual contents from PDF papers? Which tools are there? How can we mine and link data form headers and bibliography?

• SCIENTIFIC DISCOURSE CHARACTERIZATION
  How can we spot the contributions of a piece of research? Where do the authors present their future work?

• CITATION ANALYSIS
  How can citations improve our access to scientific information? Are all citations equals? How can we suggest citations?

• SCIENTIFIC DOCUMENT SUMMARIZATION
  How can we take advantage of peculiar traits of scientific documents to generate better summaries?

• CHALLENGES, DATASETS AND ARCHITECTURES
  Which datasets are available for scientific text mining? Which tasks have been proposed?

• DR. INVENTOR TEXT MINING FRAMEWORK
  Which scientific data analyses are supported? How can the framework be used in practice?

• GLOBAL CONCLUSIONS AND DISCUSSION
Outline

• How much scientific literature is there?
• How ‘open’ is scientific literature?
• Who publish scientific articles?
• How researchers search and read publications?
• Academic social networks
• Social Media in academic communication
• Text mining opportunities and challenges
Scientific literature overload

IF WE CAN FIT 140 CITATIONS PER PAGE...

...1000 PAGES PER BOOK...

A LIST OF PAPERS PUBLISHED IN 1880 WOULD FILL 100 PAGES.

BY 1920, THE LIST WOULD BE GROWING BY 500 PAGES PER YEAR.

THE 1975 SECTION WOULD FILL FOUR HUGE VOLUMES.

TODAY, WE'RE UP TO 15 VOLUMES PER YEAR—A PAGE EVERY 45 MINUTES.

BY RANDALL Munroe
Scientific Literature Overload

How much scientific literature is there?

28,100 active peer-reviewed scholarly journals in English + 6,450 non English journals
All of them are publishing 2,5 million papers a year
(more than one new article every 13 seconds)

Looking inside some citation database...

WEB OF SCIENCE™
90 million articles

Scopus
55 million articles
22,000 journals from about 5,000 publishers (in 2013 about 2 million of new articles)

crossref
80 million DOI
about 58 million refer to journal articles from 36,000 journals

Google scholar
between 100 and 160 million docs
(journal articles, books and grey literature, etc.)

PubMed
26 million publications

STM Report 2015 / citation database query
Scientific Literature Overload

How much scientific literature is there?

28,100 active peer-reviewed scholarly journals in English + 6,450 non English journals
All of them are publishing 2.5 million paper a year
(more than one new article every 13 seconds)

The number of paper published experimented an exponential growth during the last decades

The growth in number of papers is proportional to the growth in number of scientific researchers all over the world (now between 7 and 9 million, only 20% repeated authors)

STM Report 2015
How much scientific literature is there?

28,100 active peer-reviewed scholarly journals in English + 6,450 non English journals
All of them are publishing 2,5 million paper a year
(more than one new article every 13 seconds)

PubMed: from 1980 to 2003 the average growth is 2.9% per year, while from 2003 to 2013 it raised up to 6.7% per year

Web of Science: in 2000, 8,684 journals. In 2005, 9,467 journals, an increase of 9%. In 2010 11,519 journals, a further increase of 22%.
How much scientific literature is there?

...by country

The growth of scientific throughput of China: from 4.5% in 2002 to 17% currently

The citation count is dominated by USA (36%) with China in 11th place (6%), because of recent increase in scientific production
Scientific Literature Overload

Who publish scientific articles?

There are about 10,000 journal publishers globally
64% commercial publishers (including publishing for societies),
30% society publishers, 4% university publishers, 2% other publishers

Revenue:
• $10 billion in 2013 ($8 billion in 2008)
• 55% from USA, 28% Europe/Middle east, 14% Asia/Pacific, 4% Others

Employers:
• 110,000 people globally directly employed, 40% in EU (+ 20-30,000 people indirectly)

The long tail of publishing:
• the top 100 journal publisher publish 67% of all journals
• top 5 publishers are Springer, Elsevier, Wiley, Taylor&Francis (35% of all journals)
• many publishers with 1 or 2 journals
How ‘open’ is scientific literature?

The Open Access publishing model is consistently growing.

Scopus
22,000 peer-reviewed journals, 13% open-access

WEB OF SCIENCE™
8,200 peer-reviewed journals, 9% open-access

ELSEVIER
3,257 peer-reviewed journals, 17% open-access

Springer
2,500 peer-reviewed English journals, 13% open-access

DOAJ
9,237 journals, 2,330,000 articles

How ‘open’ is scientific literature?

The Open Access publishing model is consistently growing.

**Open access growth**

Before 2021, globally more than half of the papers will be published as Open Access.

PlosONE, one of the biggest Open Access journals:
- more than 34,000 articles per year (94 new articles per day)
- 2015 IF: 3.057


Scientific Literature Overload

How researchers search and read publications?

More reading, less time dedicated to each paper

**Average number of articles read by year:** about **270**
(with several variations, depending on discipline - more in medicine and science, fewer in humanities and social sciences, increased from 188 in mid-1990s)

**Reading times of an article:** about **30 minutes**
(went down from 45-50 minutes in the mid-1990s)

Clear growing importance of online literature search engines

- About 60% of article referrals of major publishers comes from one search engine, Google Scholar
- Publisher sites are accessed at article level (reduced importance of publisher site browsing)
How researchers search and read publications?

Online search and access to scientific literature

“The forced browsing of print archives may have stretched scientists and scholars to anchor findings deeply into past and present scholarship. Searching online is more efficient and following hyperlinks quickly puts researchers in touch with prevailing opinion, but this may accelerate consensus and narrow the range of findings and ideas built upon.”

Pros:
• more comprehensive searches
• more information to more extended audience

Cons:
• the articles cited tend to be more recent
• there are fewer citations
• most citations are to fewer journals and articles
• weakening ability to explore scientific literature laterally finding in other studies and disciplines information potentially relevant to our current research

Source: STM Report 2015
Scientific Literature Overload

How researchers search and read publications?

...and the coverage of search engines

About 24% of documents are freely available online with several differences across fields:

<table>
<thead>
<tr>
<th>Field</th>
<th>% of Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture Science</td>
<td>12</td>
</tr>
<tr>
<td>Arts &amp; Humanities</td>
<td>24</td>
</tr>
<tr>
<td>Biology</td>
<td>25</td>
</tr>
<tr>
<td>Chemistry</td>
<td>22</td>
</tr>
<tr>
<td>Computer Science</td>
<td>50</td>
</tr>
<tr>
<td>Economics &amp; Business</td>
<td>42</td>
</tr>
<tr>
<td>Engineering</td>
<td>12</td>
</tr>
<tr>
<td>Environmental Sciences</td>
<td>29</td>
</tr>
<tr>
<td>Geosciences</td>
<td>35</td>
</tr>
<tr>
<td>Material Science</td>
<td>12</td>
</tr>
<tr>
<td>Mathematics</td>
<td>27</td>
</tr>
<tr>
<td>Medicine</td>
<td>26</td>
</tr>
<tr>
<td>Physics</td>
<td>35</td>
</tr>
<tr>
<td>Social Science</td>
<td>19</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>43</td>
</tr>
</tbody>
</table>

Source: Khabsa & Giles (2014). The number of scholarly documents on the public web. PloS one
How researchers search and read publications?

How publishers are dealing with online scientific literature search and access?

• Improving and enriching their **online offer** and **user experience** (new tools: analytics, expertise search, etc.)

• Switching to **exclusive online publishing**:
  • currently **all STM (International Association of Scientific, Technical, and Medical Publishers) journals** can be accessible on-line – in 2003: 83%, in 2008: 96%
  • the number of **established research journals dropping their print editions** looks likely to **accelerate over the coming few years**

• Offering **enhanced article-level access** and easing the **integration of their contents** into third-party platforms by:
  • providing **enriched and linked versions of publications**
  • exposing **Open APIs**

Source: STM Report 2015
Scientific Literature Overload

Academic social networks

- ResearchGate
  more than 11 million users
  (150,000 members in August 2008, 700,000 in December 2010, one million by May 2011, and 2 million in September 2012)

- Academia.edu
  more than 44 million registered users
  (16,205,767 papers added and 1,953,015 research interests specified) Academia.edu attracts over 36 million unique visitors a month

- Mendeley
  about 4 million users
  (part of Elsevier from 2013) – 470 million documents

Why academic social networks are used?
- connecting with other researchers
- make own research more visible and follow the updates of other researchers
- like an ‘online business card’

Sources: STM Report 2015 / factsheets of academic networks
Social Media in academic communication

Social Media are experimenting an increasing adoption as complementary channel to promote and discuss scientific publications and events.

But there is still a long way to go to achieve broad Social Media adoption...

• growing impact but still limited when compared to conventional channels (in several surveys the percentage of researchers that actively use Social Media ranges from 3% to 32%)
• used mainly as complementary channels to make more visible the research than as means to interact, discuss with other users or to keep updated with new findings
• proliferation of too many potentially useful platforms to consider

Main obstacles to the use of Social Media:

• lack of clearly compelling benefits with respect to the time needed to publishing material and manage interactions
• quality and trust issues

A wider adoption of altmetrics can foster the use of Social Media in scholarly communication

Text mining opportunities and challenges

Natural Language Processing and Text Mining are starting to emerge as key technologies able to help scientists to deal with scientific literature overload.

The analysis of the structure and the semantics of full textual contents of scientific publications enables a wide range of new approaches to easily retrieve, compare and summarize scientific literature.

Citation network / data linking

Entities and relations

Scientific discourse

Content relevance

SUMMARY
Natural Language Processing and Text Mining are starting to emerge as key technologies able to help scientists to deal with scientific literature overload.

Opportunities (use cases):

- search for information scoped to specific sections of the discursive structure
- validating how scientific contents are exposed
- scientific-discourse driven summaries
- automated generation of state-of-the-art reviews
- support fast scientific literature screening, thus reducing the efforts needed for literature reviews

Challenges:

- huge, evolving amounts of data
- data collection, extraction, normalization
- error rate of automated approaches
- high variety of knowledge domains
- diversified information needs
- data often protected by copyrights
Text mining opportunities and challenges

Natural Language Processing and Text Mining are starting to emerge as key technologies able to help scientists to deal with scientific literature overload.

Access to full texts of publications: copyright issues

• Major publishers have defined their “Text and data mining policy” (limitation to text for restricted access, subject to license restriction for OA)

• Questions and answers on the modernization of EU copyright rules for the digital age - European Union, 14 September 2016:
  “The Commission proposes a new mandatory exception, which would require all Member States to permit research organizations acting in the public interest – such as universities and research institutes – to carry out text and data mining of copyright protected content to which they have lawful access, for example scientific publications they have subscribed to, without the need of a prior authorization. The exception will not apply to commercial companies.”
Scientific Literature Overload

Text mining opportunities and challenges

Natural Language Processing and Text Mining are starting to emerge as key technologies able to help scientists to deal with scientific literature overload.

Access to full texts of publications: copyright issues

Crossref
Crossref Text and Data Mining Services
http://tdmsupport.crossref.org/

Include in CrossRef metadata that describe each bibliographic entry a standard set of license information fields that clearly specify the limitations and the way to access and mine the full textual contents of papers.

• 12 publishers involved in the definition of license metadata (Elsevier, Springer, Wiley, etc.)
• common mechanism for providing automated text and data mining tools with direct links to full text on the publisher’s site
Scientific Literature Overload

**Text mining opportunities and challenges**

Natural Language Processing and Text Mining are starting to emerge as key technologies able to help scientists to deal with scientific literature overload.

This tutorial provides an overview of the core content analysis challenges and opportunities of Scientific Literature Mining showing how we can characterize and take advantage of implicit and explicit traits of scientific publications to better organize and provide access to scientific literature.

<table>
<thead>
<tr>
<th>Document Structure Analysis</th>
<th>Scientific Discourse Characterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation Analysis</td>
<td>Scientific Document Summarization</td>
</tr>
</tbody>
</table>
Outline

• Document formats: dealing with PDF
• General-purpose PDF-to-text tools
• PDF-to-text for scientific publications
• Comparing PDF-to-text for scientific publications
• Bibliographic entry parsing
• Annotated datasets
• Conclusions
Despite the many XML dialects and scientific publishing technologies proposed during the last few years, **PDF still constitutes the most widespread distribution format of scientific publications**

(80% of scientific literature is accessed as PDF documents)

**Why PDF are so popular?**

- mature technologies (1993, Adobe)
- preserved format across platform with several tools to visualize and annotate it
- easy to store and organize for off-line reading
- self contained files: capture the article in a stable, read-only form
- can include high-resolution images
- print-friendly
- can be reasonably protected without the use of dedicated servers
Despite the many XML dialects and scientific publishing technologies proposed during the last few years, PDF still constitutes the most widespread distribution format of scientific publications (80% of scientific literature is accessed as PDF documents).

Some drawbacks
- manipulation is dependent on a commercial software (even if some open-source alternative is available)
- impossible to include multimedia material / low level of interactivity with contents (internal / external hyperlinks)
- visualization not customized to the device
- difficult to extract structured textual information

Do you expect that the way they access and use articles today to change in the future? (281 responses of researchers in a variety of fields)

- Yes: 36%
- Maybe: 46%
- Not sure: 5%
- Not: 3%
PDF is a layout based data format for professional document rendering

- Each PDF document is characterized by a **body** made of a set of **objects** that are usually grouped into pages: numbers, strings, streams, arrays, dictionaries, etc.
- Each **object** in the stream is assigned a special location inside a page viewport (as well as a special size and style if applicable)
- Objects are declared in the body of the PDF files, often non sequentially
- A **cross-reference table (xref)** lists all the objects providing the file offset of each of them (optimized for reading, no need to explore the contents of the whole PDF file)
Customized approaches are necessary to extract structured textual information from the PDF of scientific publications

Title: A Comparison of Layout based Bibliographic Metadata Extraction Techniques
Author 1 Name: Michael Grantizer
Author 1 Affiliation: University of Passau, Germany
Author 1 Email: Michael.Grantizer@uni-passau.de
Abstract: Social research networks such as Mendeley and CiteULike offer various services for collaboratively managing bibliographic metadata...

The quality of scientific text mining often depends in the first place to the quality of the extraction of semi-structured textual contents from the original PDF file

- robust with respect to different document layouts
  a random sample of 125,000 publications from PubMed contains articles of 500 publishers, each one with its own layout and style
- covering and customized to the wide set of structural elements of scientific articles
General purpose PDF-to-text software

Convert PDF files to plain text, XML / HTML files with some layout information

- **PDFBox**
  - Java library
  - Apache Project (actively maintained)
  - 2.0.3 released on 17/9/2016

- **pdf2xml**
  - C++ code
  - Freeware, based on xpdf
  - 2.1 released on 14/6/2014

- **Poppler**
  - C++ code
  - Open source, based on xpdf
  - 0.49.0 released on 15/11/2016

  Used by GIMP, Okular, Pdf2HTMLex, etc.

- **JPedal**
  - Commercial
  - Specialized to extract tables, word lists and other elements

...also Adobe Acrobat, iText, etc.

Try and compare these tools and others at: [http://backingdata.org/pdfconv/](http://backingdata.org/pdfconv/)
General purpose PDF-to-text software

Example of XML output generated by pdf22xml
General purpose PDF-to-text software

pdf2htmlEX

Open source project (GPL v3) of layout-preserving PDF to HTML converter
(C++ mainly, based on Poppler to process PDF files)

Each PDF file is converted in an HTML file made of a set of DIV elements properly positioned inside the page viewport

---

**Trace-based Just-in-Time Type Specialization for Dynamic Languages**

_andreas.gol@Mozilla.com, brendan.eich@mozilla.com, mike.shaver@mozilla.com, david.anderson@mozilla.com, david.mandelin@mozilla.com, mohammad.r.haghighat@intel.com, blake.kaplan@mozilla.com, graydon.houre@mozilla.com, boris.zbarsky@mozilla.com, jason.orendorff@mozilla.com, jesse.ruderman@mozilla.com, edwin.smith@mozilla.com, rick.reitmaier@mozilla.com, michael.bebenita@mozilla.com, mason.chang@mozilla.com, michael.franz@mozilla.com_

Mozilla Corporation

---

**HTML source**

```html
...<div class="t x1 h1 y1 ff1 fs0 fc0 ws0">Languages</div>
...
```

---

**CSS CLASSES TO DEFINE THE LAYOUT**

```css
.x1 { left: 441.233569px; }
.y1 { bottom: 1139.024816px; }
.h1 { height: 49.637990px; }
.ff1 { font-family: ff1; line-height: 0.898000; }
.fs0 { font-size: 71.731200px; }
.fc0 { color: rgb(0,0,0); }
.ws0 { word-spacing: 0.000000px; }
.t { position: absolute; white-space: pre; }
```
PDF-to-text for scientific publications

• Integrating a **general purpose PDF-to-text converter** and post processing its output

• Implementing **customized approaches to identify structural elements proper of scientific publications** like: title, authors and affiliations, abstract, section heading and contents, figures, tables, formulas, bibliographic entries, in-line citation markers, etc.

• Robust to varied publishing styles
PDF-to-text for scientific publications

**CERMINE**
Content ExtRactor and MINEr

- Java, open-source (GitHub) - libSVM
- PDF analysis: based on both layout features and contents
- Output: JATS XML

**PDF-TO-XML CONVERSION**

- **METADATA ZONE CLASSIFICATION**
  
  **Algorithm**: SVM
  **Classes**: title, author, affiliation, editor, correspondence, type, abstract, keywords, bib_info, dates

- **HIGH-LEVEL ZONE CLASSIFICATION**
  
  **Algorithm**: SVM
  **Classes**: metadata, references, body
  **Feats**: geometric, sequential, lexical, heuristics on content

- **EXTRACTION OF HIERARCHY OF BASIC STRUCTURAL ELEMENTS**

  Split words into chars (height, width, x, y), then apply Docsum:
  - **hierarchical grouping of elements** relying on: nearest neighbor pairs of chars, averages distances among chars and lines, line parallelness and overlap, etc.
  - hierarchy of structural elements: chars, words, lines, zones, pages

- **BIB. ENTRIES IDENTIFICATION**
  
  **Algorithm**: CRF
  **Classes**: firstname, surname, title, source, volume, infopage, year, etc.
  **Feats**: 42 like special chars, match in dict. lists, etc.

- **BIB. ENTRIES PARSING**
  
  **Algorithm**: kMeans (k=2) to group reference lines into first line of bib. entry / other
  **Feats**: line length, indentation, space between lines, etc.
Document Structure Analysis

PDF-to-text for scientific publications

CERMINE
Content ExtRactor and MINEr

http://cermine.ceon.pl/
https://github.com/CeON/CERMINE

TRAINING AND EVALUATION OF SVM / CRF

Zone classifier (high-level and metadata) trained and evaluated on the GROTOAP2 dataset:
• 2,651 document from PubMed available as PDF + JATS XML
• for each PDF + JATS XML pair of files:
  • PDF processed by CERMINE
  • JATS XML annotations used to label the zones identified by CERMINE (text sequence alignment algorithm)
• analyzing a sample of the subset of the transferred annotations, a set of heuristic rules to improve the quality of annotation transfer is developed and applied to each document

Citation parser trained and evaluated on the three datasets: 4,000 parsed citations: 2,000 from CiteSeer and Cora-ref and 2,000 from 1991 different PMC documents

RESULTS:
• Citation parsing F-score: 93.3%
• Metadata and Bibliography extraction F-score (47,983 PDF + metadata records): 77.5%

PDF-to-text for scientific publications

GROBID

- Java (with JNI call to native CRF libraries: CRF++ or Wapiti) open-source (GitHub)
- PDF analysis with CRF exploiting both layout features and contents
- Output: TEI XML

HIERARCHY OF CRF SEQUENCE TAGGERS

- position information (begin/end of line, in the doc.)
- lexical information (vocabulary, large gazetteers)
- layout information (font size, block, etc.)

Document segmentation hierarchy

- 9 CRF models for full texts
- 14 intermediary labels in total
- 55 final labels

Document Structure Analysis

PDF-to-text for scientific publications

GROBID

https://grobid.readthedocs.io/
https://github.com/kermitt2/grobid

TRAINING AND EVALUATION OF CRF

• Each model with its own set of features, specialized to tag certain fields
• Training sets specific to each model included in the software (see table)
• Trainer framework to generate and manually validate new training examples from a collection of PDF files
• Best performing header metadata extraction tool over 7 (Lipinski et al. 2013)

<table>
<thead>
<tr>
<th>Model</th>
<th>Nº training examples</th>
<th>Exploit layout</th>
</tr>
</thead>
<tbody>
<tr>
<td>segmentation</td>
<td>121</td>
<td>x</td>
</tr>
<tr>
<td>header</td>
<td>3971</td>
<td>x</td>
</tr>
<tr>
<td>affiliation-address</td>
<td>1064</td>
<td></td>
</tr>
<tr>
<td>names (header)</td>
<td>1297</td>
<td></td>
</tr>
<tr>
<td>names (citation)</td>
<td>253</td>
<td></td>
</tr>
<tr>
<td>date</td>
<td>619</td>
<td></td>
</tr>
<tr>
<td>reference-segmenter</td>
<td>17</td>
<td>x</td>
</tr>
<tr>
<td>citation</td>
<td>4150</td>
<td></td>
</tr>
<tr>
<td>fulltext (body)</td>
<td>8 (+13 abstracts)</td>
<td>x</td>
</tr>
</tbody>
</table>

A customized versions of GROBID is exploited by:

Extraction of bibliographic entries and matching against internal DB
• about 300,000 PDF processed monthly (16 nodes Hadoop cluster)
• failure rate of 1% of user uploaded PDF

Document Structure Analysis

PDF-to-text for scientific publications

http://pdfx.cs.man.ac.uk/

- Online Web Service (Max 5Mb)
- PDF analysis is rule-based, relying on both layout features and contents
- Output: JATS compliant XML files

Two steps PDF analysis:

STEP 1: a geometrical model of the textual contents and the layout of the information contained in the PDF is build:
  - each word described is by orientation, position, font, etc.
  - global document stats: most frequent font size and style, average line spacing and font spacing, etc.
  - neighbor words sharing similar text features are merged into blocks

STEP 2: based on the layout features previously spotted, a set of rules is exploited to merge blocks into regions and iteratively identify 18 elements
Inside the document (also on the basis of the surrounding elements)

Document Structure Analysis

**PDF-to-text for scientific publications**

[pdfx v1.9](http://pdfx.cs.man.ac.uk/)

**EVALUATION**

- 50,000 PDF + XML articles published by Elsevier in 2008
- 1,943 PDF + XML articles published in the PMC Open Access Subset in 2011 each one from a different journal

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>h3</th>
<th>table</th>
<th>h2</th>
<th>fig_tbl_ref</th>
<th>abstract</th>
<th>caption</th>
<th>author</th>
<th>citation</th>
<th>bib_item</th>
<th>h1</th>
<th>email</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elsevier</td>
<td>50000</td>
<td>83.35</td>
<td>28.78</td>
<td>82.03</td>
<td>89.1</td>
<td>62.01</td>
<td>82.86</td>
<td>94.63</td>
<td>75.46</td>
<td>86.08</td>
<td>90.5</td>
<td>97.61</td>
<td>96.7</td>
</tr>
<tr>
<td>PMC_sample</td>
<td>1943</td>
<td>6.05</td>
<td>13.27</td>
<td>27.19</td>
<td>27.52</td>
<td>32.41</td>
<td>54.53</td>
<td>61.65</td>
<td>63.10</td>
<td>74.03</td>
<td>77.45</td>
<td>79.67</td>
<td>85.42</td>
</tr>
</tbody>
</table>

F-score per class – 0.95 similarity threshold between extracted and original textual contents of each field

- Elsevier dataset more curated
- PMC dataset suffers the high variation in style due to the presence of 1,943 articles each one from a different journal
- *tables* are with both datasets difficult to identify, while *title* and *email* are the easiest to spot

PDF-to-text for scientific publications

- SectLabel

https://github.com/knmnyn/ParsCit/tree/master/bin/sectLabel

- Perl and Ruby (CRF++), open-source (GitHub), process Omnipage output
- PDF analysis relying on both layout features and contents

**EVALUATION**
Adding to textual / content features also layout features improves the F-score of about 10 points (up to 84%) and is particularly beneficial for sections like metadata, captions, hierarchical headers

**ALL LINES LABELING**
- **Algorithm**: CRF
- **Classes (26)**: address, affiliation, author, bodyText, categories, construct, copyright, email, equation, figure, figureCaption, footnote, keywords, listItem, note, page, reference, sectionHeader, subsectionHeader, subsubsectionHeader, table, tableCaption, title
- **Line feats**: location, number, punctuation, length, format / layout, differences in format / layout with previous and following lines

**HEADER LINE LABELING**
- **Algorithm**: CRF, applied only to lines classified as Header in the previous step
- **Classes (13)**: abstract, categories, general terms, keywords, introduction, background, related work, methodology, evaluation, discussions, conclusions, acknowledgements, references
- **Line feats**: location, number, punctuation, length, format / layout, differences in format / layout with previous and following lines

Document Structure Analysis

PDF-to-text for scientific publications

lapdftext

https://github.com/BMKEG/ lapdftextProject

• Java, open-source (GitHub)
• PDF analysis is rule-based, relying on both layout features and contents
• Output: JATS XML

1. Word blocks merged into text blocks by relying on page-level and document-level features like distance between words and lines, font sizes and weight, etc.

2. Rules exploited to assign to each text blocks a specific class among: title, abstract, heading, sub-heading, references, etc.

3. Other rules are used to define the reading order of classified text blocks and eventually merge together contiguous text blocks belonging to the same class.

How header metadata are extracted from a PDF document?

Initial approach:
PDF to text by means of PDFnet software (commercial) then apply **iterative multi-step SVM classifier (RBF kernel)** as in:


1) **SVM that independently classifies each header line** with respect to textual features: position, number of words, number of capitalized words, % of words in specific dictionaries, % of words occurring in a specific class of training data

2) **Contextual iterative classification by SVM** where each header line is described by the feature set at step 1 and the class assigned to the previous L and next N header lines – **stop condition**: label assignments to header lines changes less than a predefined threshold with respect to previous iteration

3) Proper heuristics to **segment lines with multiple authors**
How header metadata are extracted from a PDF document?

Later approach:

- Trained on a large set of papers
- **Fields**: title, authors, DOI, publication, volume, issue, year, page, ranges

Evaluation:

- **Dataset**: 26,000 PDFs with perfect metadata record in Mendeley Catalogue
  - 2.4% couldn’t be converted to XML by pdftoxml
  - 83.9% can be processed extracting perfect metadata records: authors, title, year, and publication venue (e.g. journal, conference, magazine)

“If you drop 10 PDFs into your Mendeley Library then, on average, you’ll get perfect, citable metadata for 8-9 of them.”

https://mendeleyapi.wordpress.com/2014/10/15/pdf-extraction-gets-a-boost-with-our-new-api-service/
https://krisjack.wordpress.com/2015/03/12/how-well-does-mendeleys-metadata-extraction-work/
Comparing PDF-to-text for scientific publications


Dataset:
1,153 random PDF articles from arXiv together with their metadata (title, authors, year, abstract) dealing with Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

Evaluation:
A_{100}: 100 randomly selected articles and manual evaluation: 1 perfect match, 0.5 accent or ligature issues, 0.25 partial match, 0 no match
B_{100}: 100 randomly selected articles and automated evaluation: Levenshtein distance normalized by the length of the reference value for the field
B_{1153}: whole dataset and automated evaluation: Levenshtein distance normalized by the length of the reference value for the field

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Title</th>
<th>Authors</th>
<th>Authors' last names</th>
<th>Abstract</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROBID</td>
<td>A_{100}</td>
<td>B_{100}</td>
<td>B_{1153}</td>
<td>A_{100}</td>
<td>B_{100}</td>
</tr>
<tr>
<td>Mendeley Desktop</td>
<td>N/A</td>
<td>0.92</td>
<td>0.92</td>
<td>N/A</td>
<td>0.83</td>
</tr>
<tr>
<td>ParsCit</td>
<td>SectLabel</td>
<td>0.59</td>
<td>0.52</td>
<td>0.54</td>
<td>0.47</td>
</tr>
<tr>
<td>PDFSSA4MET</td>
<td>0.13</td>
<td>0.21</td>
<td>0.18</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>PDFMeat</td>
<td>0.60</td>
<td>N/A</td>
<td>N/A</td>
<td>0.6</td>
<td>N/A</td>
</tr>
<tr>
<td>SciPlore Xtract</td>
<td>0.76</td>
<td>0.81</td>
<td>0.78</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SVMHeaderParse</td>
<td>0.50</td>
<td>0.57</td>
<td>0.61</td>
<td>0.64</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Accuracy values
Comparing PDF-to-text for scientific publications


Dataset:
1,943 pairs of PDF + JATS XML documents retrieved from PubMed Open Access Subset

Evaluation:
metadata extraction of CERMINE and other 4 similar tools (exact match)

<table>
<thead>
<tr>
<th></th>
<th>CERMINE</th>
<th>PDFX</th>
<th>GROBID</th>
<th>ParsCit</th>
<th>Pdf-extract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>95.5</td>
<td>85.7</td>
<td>82.5</td>
<td>34.1</td>
<td>49.4</td>
</tr>
<tr>
<td></td>
<td>93.4</td>
<td>84.7</td>
<td>77.4</td>
<td>39.6</td>
<td>49.4</td>
</tr>
<tr>
<td></td>
<td><strong>94.5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>90.2</td>
<td>71.2</td>
<td>85.9</td>
<td>57.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>89.0</td>
<td>71.5</td>
<td><strong>90.5</strong></td>
<td>48.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>89.6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affiliations</td>
<td>88.2</td>
<td>-</td>
<td>90.8</td>
<td>72.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>83.1</td>
<td>-</td>
<td>51.8</td>
<td>44.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>85.6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email addresses</td>
<td>51.7</td>
<td>53.0</td>
<td>26.9</td>
<td>28.8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>42.6</td>
<td><strong>73.6</strong></td>
<td>7.8</td>
<td>36.2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>46.7</td>
<td><strong>61.6</strong></td>
<td>12.1</td>
<td>32.1</td>
<td>-</td>
</tr>
<tr>
<td>Abstract</td>
<td>82.8</td>
<td>71.1</td>
<td>70.4</td>
<td>47.7</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>79.9</td>
<td>66.7</td>
<td>67.7</td>
<td>61.3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>81.3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td>89.9</td>
<td>-</td>
<td><strong>94.2</strong></td>
<td>15.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>63.5</td>
<td>-</td>
<td>44.2</td>
<td>3.0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>74.4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal</td>
<td>80.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>73.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>76.6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volume</td>
<td>93.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>83.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>87.8</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Issue</td>
<td>53.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>28.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>37.1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pages</td>
<td>87.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>80.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>83.5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>96.3</td>
<td>-</td>
<td>95.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>95.0</td>
<td>-</td>
<td>40.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>95.6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td>98.2</td>
<td>-</td>
<td>99.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>75.0</td>
<td>-</td>
<td>65.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td><strong>85.1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>References</td>
<td>96.1</td>
<td>91.3</td>
<td>79.7</td>
<td>81.2</td>
<td>80.4</td>
</tr>
<tr>
<td></td>
<td>89.8</td>
<td>88.9</td>
<td>66.7</td>
<td>71.8</td>
<td>57.5</td>
</tr>
<tr>
<td></td>
<td><strong>92.8</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In every cell there is precision, recall and F-score value
Bibliographic entry parsing

Tools:
• most of them based on: rules or sequence taggers like HMM and CRF
• some example:
  ➢ **FreeCite**: CRF++ library, CORA dataset, open-source (ruby), Web API
    [http://freecite.library.brown.edu/](http://freecite.library.brown.edu/)
  ➢ **ParsCit**: CRF++ library, open source (perl and C++)
    [https://github.com/knmnyn/ParsCit](https://github.com/knmnyn/ParsCit)

Web API:
• Web services that match against a citation database:
  ➢ **CrossRef Metadata Search API**: find metadata by DOI or by bibliographic entry string
    [http://search.crossref.org/help/api](http://search.crossref.org/help/api)
  ➢ **Bibsonomy REST API – search posts by string**:
    [https://bitbucket.org/bibsonomy/bibsonomy/wiki/documentation/api/REST%20API](https://bitbucket.org/bibsonomy/bibsonomy/wiki/documentation/api/REST%20API)
    [https://www.bibsonomy.org/api/posts?resourcetype=bookmark&search=SemKey](https://www.bibsonomy.org/api/posts?resourcetype=bookmark&search=SemKey)
Annotated datasets

CORA Field Extraction dataset

https://people.cs.umass.edu/~mccallum/data/cora-ie.tar.gz


- 500 tagged references: author, title, journal, volume, pages, date
- 937 headers: title, author, affiliation, address, email, abstract, keywords

FluXcim Citation dataset

https://github.com/knmnyn/ParsCit/blob/master/doc/flux-cim-cs.tagged.txt


- 300 citation strings randomly from ACM Digital Library - CORA format

UMASS Citation dataset

http://www.iesl.cs.umass.edu/data/umasscitationfield


- from arXiv papers in physics, mathematics, computer science and quantitative biology
- 1,800 citations hierarchically labeled as:
  - ref-markers
  - author → first, middle, last and affix
  - title
  - venue → publisher, note, web, institution, department, etc.
  - date → year and month
  - reference-id

![Diagram of citation field extraction]
Document Structure Analysis

Annotated datasets

**CiteSeer Citation dataset**
https://github.com/knmnyn/ParsCit/blob/master/doc/citeseerx.tagged.txt


  - 200 tagged references: author, title, journal, volume, pages, date

**GROTOAP2 - GROund Truth for Open Access Publications**
http://cermine.ceon.pl/grotoap2/


  - 13,210 ground truth files in TrueViz XML format (1,640,973 zones in total) – each one corresponding to a PDF + JATS XML of the Open Access Subset of PubMed Central
  - thanks to TrueViz, each file is represented as a hierarchy of structural elements:
    - a list of pages
    - each page contains a list of zones
    - each zone contains a list of lines
    - each line contains a list of words
    - and finally each word contains a list of characters
  
Structural elements have: text content, position on the page and dimensions. Also the natural reading order for all structure elements is specified.

Each zone has labels (imported from PubMed) describing the role in the document are assigned to zones. There are 22 labels including: abstract, acknowledgments, affiliation, author, bib_info, body_content, conflict_statement, copyright, dates, editor, equation, etc.
Conclusions

• A precise extraction of structured textual contents from the PDF of scientific publications is essential to enable any further text processing of their contents

• Several PDF-to-text conversion tools are available, both general purpose and customized to scientific publications

• Such tools usually rely on both layout and textual features of scientific publications and are rule-based or rely on supervised machine learning approaches

• A rich set of annotated corpora is freely available for further experimentation
SCIENTIFIC DISCOURSE
CHARACTERIZATION
Outline

• What is scientific discourse?
• Scientific discourse characterization
  – Annotation procedures and annotated corpora
  – Automated annotation of scientific texts
• Overview of available datasets
• Conclusions
What is scientific discourse?

Scientific discourse concerns the characterization of how content is presented, discussed and motivated in scientific literature.

The most effective sources of supervision for training statistical parsers are treebanks. Unfortunately, treebanks are expensive, time-consuming to create, and not available for most domains.

BACKGROUND

The most effective sources of supervision for training statistical parsers are treebanks.

CONTRAST

Unfortunately, treebanks are expensive, time-consuming to create, and not available for most domains.

APPROACH

OUTCOME

The most effective sources of supervision for training statistical parsers are treebanks.
Scientific discourse concerns the characterization of how content is presented, discussed and motivated in scientific literature.

Why making scientific discourse explicit?

*Provide new dimensions to drive the automated analysis of scientific publications*

- ease the interpretation of the information flow
- contextualize contents and characterize their connections with related pieces of research
- discover relevant aspects, novelties and future directions
- support tasks like targeted information extraction, content retrieval and summarization
- assess the quality of content exposition
Scientific discourse characterization

Steps towards the characterization and automated annotation of scientific discourse

Annotation procedure and annotated corpus
- definition of annotation schema
- corpus annotation:
  - corpus content selection
  - annotation guidelines and procedure
  - annotation results

Automated annotation of scientific texts
- algorithmic approach
- feature engineering

BACKGROUND
- APPROACH
- BACKGROUND
- FUTURE WORK
Scientific discourse characterization

Two main approaches to the rhetorical analysis of a text:

**Rhetorical Structure Theory:** relations between clauses or larger text segments

**Zone Analysis:** characterization of the global rhetorical status of each text unit (sentence)

The most effective sources of supervision for training statistical parsers are treebanks.

Unfortunately, treebanks are expensive, time-consuming to create, and not available for most domains.
Rhetorical Structure Theory (Mann & Thompson)

Coherent texts consist of **minimal units**, which are linked to each other, recursively, through rhetorical relations thus generating a tree-like representation of a text.

Many neo-pagan religions, such as Wicca, use aspects of ancient Greek religions in their practice; Hellenic polytheism instead focuses exclusively on the old religions, as far as the fragmentary nature of the surviving source material allows.
Scientific Discourse Characterization

Scientific discourse characterization

Annotation procedure and annotated corpus

Rhetorical Structure Theory (Mann & Thompson)

Coherent texts consist of **minimal units**, which are linked to each other, recursively, through rhetorical relations thus generating a tree-like representation of a text

- **23 relations** (symmetric and not; several extensions proposed)
- the text is represented by a **recursive tree structure** (the most relevant minimal units are usually placed on the top)

Among others, exploited to:
- Check text coherence
- Natural Language Generation
- Corpus analysis and study of discourse phenomena
- Text summarization

Many neo-pagan religions, such as Wicca, use aspects of ancient Greek religions in their practice; Hellenic polytheism instead focuses exclusively on the old religions, as far as the fragmentary nature of the surviving source material allows.
Several annotations schemes and procedures have been proposed to characterize text units with respect to:

- the **type and complexity of the discourse elements** identified
- the **type of text units** to which the discourse is applied (sentences, segments of sentences, specific relations or events occurring in these sentences)

Knowledge Claim discourse Model and Argumentative Zoning

Core Scientific Concepts

IMRAD structure

Dr. Inventor Scientific Discourse Schema
Scientific discourse develops throughout a set of ‘rhetorical moves’ that are explicit statements, referred to as Knowledge Claims, useful to characterize and justify the contributions of a specific piece of work.

**EXAMPLES OF RHETORICAL MOVES**

- **Properties of research space**
- **Properties of new solution (US)**
- **Properties of existing solution (THEM)**
- **Relationship between existing and new solution (US and THEM)**

- Open domain word sense disambiguation presents several interesting challenges both semantic and computational.

- The proposed methodologies solves the issues related with the high computational cost of knowledge analysis.

- The method proposed by Ray et. al., 2010 stressed the importance of correctly dealing with semantic draft.

- Our solution improves the previous state-of-the-art method (Gil et al., 2012) by exploiting a new set of data sources.
Annotation categories are defined on the basis of who owns the knowledge claim.

AZ Corpus

- **80 conference articles** in computational linguistics
- **12,188 sentences** assigned to one of 7 Categories
- Avg. annotator agreement K: **0.71**

Online at (SciXML format):

http://www.cl.cam.ac.uk/~sht25/AZ_corpus.html
Scientific Discourse Characterization

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: Argumentative Zoning cross-domain validity

Test if non-expert humans can annotate the sentences of text from different domains (Computational Linguistics and Chemistry) with respect to an extended version of the Argumentative Zoning Schema

Argumentative Zoning II Schema (to model typical Chemistry argumentation)

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIM</td>
<td>Statement of specific research goal or hypothesis of current paper</td>
<td>OWN_CONC</td>
<td>Findings, conclusions (non-measurable) of own work</td>
</tr>
<tr>
<td>NOV_ADV</td>
<td>Novelty or advantage of own approach</td>
<td>CODI</td>
<td>Comparison, contrast, difference to other solution (neutral)</td>
</tr>
<tr>
<td>CO_GRO</td>
<td>No knowledge claim is raised (or knowledge claim not significant for the paper)</td>
<td>GAP_WEAK</td>
<td>Lack of solution in field, problem with other solutions</td>
</tr>
<tr>
<td>OTHER</td>
<td>Knowledge claim (significant for paper) held by somebody else. Neutral description</td>
<td>ANTISUPP</td>
<td>Clash with somebody else’s results or theory; superiority of own work</td>
</tr>
<tr>
<td>PREV_OWN</td>
<td>Knowledge claim (significantly held by authors in a previous paper. Neutral description.</td>
<td>SUPPORT</td>
<td>Other work supports current work or is supported by current work</td>
</tr>
<tr>
<td>OWN_MTHD</td>
<td>New knowledge claim, own work: methods</td>
<td>USE</td>
<td>Other work is used in own work</td>
</tr>
<tr>
<td>OWN_FAIL</td>
<td>A solution/method/experiment in the paper that did not work</td>
<td>FUT</td>
<td>Statements/suggestions about future work (own or general)</td>
</tr>
<tr>
<td>OWN_RES</td>
<td>Measurable/objective outcome of own work</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data:
- 30 Chemistry papers
- 9 Computational Linguistics papers (CL)

Annotation:
3 annotators experts in CL with different levels of expertise in Chemistry → chemistry intro
Test if non-expert humans can annotate the sentences of text from different domains (Computational Linguistics and Chemistry) with respect to an extended version of the Argumentative Zoning Schema.

**Argumentative Zoning II Schema** (to model typical Chemistry argumentation)

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AZ Aim</strong></td>
<td>Aim</td>
<td><strong>AZ Own</strong></td>
<td>Own.Conc: Findings, conclusions (non-measurable) of own work</td>
</tr>
<tr>
<td><strong>AZ Background</strong></td>
<td>Co_Gro</td>
<td><strong>AZ Contrast</strong></td>
<td>Codi: Comparison, contrast, difference to other solution (neutral)</td>
</tr>
<tr>
<td><strong>AZ Other</strong></td>
<td>Other</td>
<td><strong>AZ Contrast</strong></td>
<td>Gap_Weak: Lack of solution in field, problem with other solutions</td>
</tr>
<tr>
<td><strong>AZ Other</strong></td>
<td>PrevOwn</td>
<td><strong>AZ Basis</strong></td>
<td>Antisupp: Clash with somebody else’s results or theory; superiority of own work</td>
</tr>
<tr>
<td><strong>AZ Own</strong></td>
<td>Own_Method</td>
<td></td>
<td>Support: Other work supports current work or is supported by current work</td>
</tr>
<tr>
<td><strong>AZ Own</strong></td>
<td>Own_Fail</td>
<td></td>
<td>Use: Other work is used in own work</td>
</tr>
<tr>
<td><strong>AZ Own</strong></td>
<td>Own_Res</td>
<td></td>
<td>Fut: Statements/suggestions about future work (own or general)</td>
</tr>
</tbody>
</table>
Scientific Discourse Characterization

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning


**Classifier**: Naïve Bayes

**Sentence features**:

- **Structural**:
  - Absolute sentence location
  - Position of sentence within section and paragraph
  - Type of headline of current section (15 prototypical types)
  - Words shared with title or headlines
  - Significant words (sentences that contain one of the 18 highest TF*IDF words)
  - contain self-citation

- **Sentence-scoped**:
  - Verb (voice, tense, modal)
  - Contain citation
  - Most probable previous sentence category
  - Meta-discourse expression (formulaic expressions, type of agent, type of action)

<table>
<thead>
<tr>
<th>AIM</th>
<th>CONTR.</th>
<th>TEXTUAL</th>
<th>OWN</th>
<th>BACKG.</th>
<th>BASIS</th>
<th>OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>52</td>
<td>44</td>
<td>65</td>
<td>26</td>
<td>34</td>
<td>20</td>
<td>61</td>
</tr>
<tr>
<td>57</td>
<td>66</td>
<td>88</td>
<td>45</td>
<td>40</td>
<td>50</td>
<td>38</td>
</tr>
<tr>
<td>37</td>
<td>40</td>
<td>44</td>
<td>44</td>
<td>52</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>System</td>
<td>52</td>
<td>44</td>
<td>65</td>
<td>26</td>
<td>34</td>
<td>20</td>
</tr>
</tbody>
</table>
Scientific discourse characterization

Automated annotation of scientific texts

**Zone Analysis: Argumentative Zoning**

Argumentative zoning sentence classifier robust with respect to noisy input: **plain text** or textual input generated from PDF to text conversion or OCR

Explicit structure (SciXML)

```
<TITLE>Paper title</TITLE>
<HEDER>Section title</HEADER>
<S>First sentence of the paper.</S>
<S>Second sentence of the paper.</S>
....
```

Plain textual contents

```
Paper title
Section title
First sentence of the paper. Second sentence of the paper.
....
```

Classification without using structural features: Absolute sentence location / Position of sentence within section and paragraph / Type of headline of current section / Words shared with title or headlines / Significant words (TF*IDF) / Is self-citation
Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Argumentative zoning sentence classifier robust with respect to noisy input: plain text or textual input generated from PDF to text conversion or OCR

Classifier: Maximum entropy (automatically spotted and POS tagged sentences)

Sentence-scoped features:
- Normalized number of sentences from the beginning
- Overlap with first 100 words of text
- Verb (voice, tense, modal)
- Contain citation, is self-citation
- Set reduced to agent type
- Raw tokens, bigrams, trigrams

The agreement with the Gold Standard, even with noisy input data (PDF to text extractor, automatic sentence and paragraph identification and POS tagging) is still respectable and the classifier is still robust and fast to execute.
Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

The linguistic constructs that are used to express the rhetorical functions in a paper are **independent from the topic**

The problem of ___________________________ has received a lot of attention because of its relevance to _________________________________. ___ proposed an approach based on _________________________________. In this paper we present a method to _________________________________. We demonstrate the empirical effectiveness of our method reporting experiment on _________________________________.

**Topic-independent template for abstracts of NLP papers**
Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Two language models can be composed by a binary-valued latent variable to generate the words of a paper:

- **LDA topic model**: to generate the topic dependent words of a document
- **Word distribution of a rhetorical zone**: to represent transition probabilities across rhetorical categories of sentences. A Markov model is used since the probability of a zone is dependent on the zone of the previous sentence.

The problem of Word Sense Disambiguation has received a lot of attention because of its relevance to the correct interpretation and integration of textual contents. We proposed an approach based on knowledge resources built with unsupervised approaches from a corpus. In this paper we present a method to extend semantic networks to improve their effectiveness on Word Sense Disambiguation. We demonstrate the empirical effectiveness of our method reporting experiment on a wide collection of sense annotated corpora.
Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Given a *number collection of documents*, a *number of topics* and a *number of rhetorical zones* to discover, this unsupervised approach assign each sentence to:
• a distribution of topics $\rightarrow$ most likely topic
• a distribution of rhetorical zones $\rightarrow$ most likely zone

The problem of Word Sense Disambiguation has received a lot of attention because of its relevance to the correct interpretation and integration of textual contents.
How good is this approach to cluster sentences into rhetorical zones?

**Dataset**: 1000 abstracts annotated with Argumentative Zoning

**Zone clustering approaches**:
- Boilerplate-LDA (presented in the paper)
- Boilerplate-LDA with probability of zone transition independent from adjacent sentences (no Markov model for zone transition)
- Boilerplate-LDA without topics
- **kMeans** (FEATURES: tf-idf-transformed lexical frequencies, part-of-speech tags and a location feature computed by dividing the abstract into 5 bins)

Compared with Gold Standard sentence clustering into zones

**Results**:
Boilerplate-LDA (presented in the paper) generates clusters of sentences that are more consistent with Gold Standard clusters
Can we use learned zones as features to improve supervised classification?

Dataset: 1000 abstracts annotated with Argumentative Zoning

Classification approaches: Logistic Regression with history feature and CRF

- Base features: tf-idf-transformed lexical frequencies, part-of-speech tags and a location feature computed by dividing the abstract into 5 bins extended with:
  1. Boilerplate-LDA zone feature (index of the zone from 1 to 10)
  2. Topics that are assigned to the words of a sentence by LDA are set to true (one feature per topic)
  3. Only topic that is assigned with more frequency set to true (one feature per topic)

Results:
Performance of a Logistic Regression and CRF classifier improves with the addition of zone features (item 1, helps to identify the rhetorical zone) and not with the addition of topic features
Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning


- sentence features: unigram, bigram, section counter, location inside section and paragraph and length
- improvement of sentence classification performance on Argumentative Zoning corpus by using a **maximum entropy classifier**
- by using an **HMM** with only unigrams and bigrams the classification accuracy improvement is relevant up to an history of the four previous decisions


- **SciPo**: tools that applies a set of rules to evaluate the coherence of scientific abstracts of novices on the basis of their rhetorical structure spotted by AZ classifier
- **Argumentative Zoning schema ported to scientific abstract in Portuguese**: the category OWN divided into Methodology, Results and Conclusion
- Corpus of **52 abstracts annotated**
- Automated classification experiments with Teufel’s features ported to Portuguese: Classifier: Naïve Bayes (13-folds cross validation) accuracy 74%, K with gold standard 0.65
Scientific Discourse Characterization

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning


- 20 online articles taken from major biology journals annotated in order to develop and refine the annotation schema on the bases of Teufel’s Argumentative Zoning
- extended modified version of AZ Schema to include:
  - a finer grained classification of the author’s own work
  - an explicit relation between the data presented and the findings


- adaptation of Argumentative Zoning to the legal domain
- unlike scientific texts, the fundamental communicative purpose of a judgment is to legitimise a decision, by showing that it derives, by a legitimate process, from authoritative sources of law
- schema categories: FACTS, PROCEEDING, BACKGROUND, FRAMING, DISPOSAL, TEXTUAL, others
Scientific discourse characterization

Zone Analysis: Core Scientific Concepts

A paper is a human readable representation of a scientific investigation: a scientific discourse annotation schema should point out the components of the scientific investigation.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis</td>
<td>An unconfirmed statement which is a stepping stone of the investigation</td>
</tr>
<tr>
<td>Motivation</td>
<td>The reason behind the investigation</td>
</tr>
<tr>
<td>Background</td>
<td>Generally expected background knowledge and previous work</td>
</tr>
<tr>
<td>Goal</td>
<td>A target state of the investigation where intended discoveries are made</td>
</tr>
<tr>
<td>Object</td>
<td>An entity which is the product or main theme of the investigation (advantage/disadvantage)</td>
</tr>
<tr>
<td>Method</td>
<td>Means by which the authors seek to achieve the goal of the investigation (old/new – advantage/disadvantage)</td>
</tr>
<tr>
<td>Experiment</td>
<td>An experimental method</td>
</tr>
<tr>
<td>Model</td>
<td>A statement about a theoretical model or framework</td>
</tr>
<tr>
<td>Observation</td>
<td>The data / phenomena recorded in an investigation</td>
</tr>
<tr>
<td>Result</td>
<td>Factual statements about the outputs, interpretation of an observation</td>
</tr>
<tr>
<td>Conclusion</td>
<td>Statements inferred from observations and results</td>
</tr>
</tbody>
</table>

ART Corpus

- 265 papers from the domains of chemistry and biochemistry
- 39,915 sentences
- Avg. annotator agreement K: 0.55

Scientific Discourse Characterization

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Core Scientific Concepts


Classifiers: SVM (linear), CRF

Sentence features:

• Structural:
  • Absolute sentence location
  • Section ID (incremental integer, up to 10)
  • Length and position of sentence within section and paragraph
  • Type of headline of current section (16 types of prototypical headers)

• Sentence-scoped:
  • No citations, one citation, +1 citation
  • Category of previous sentence (not CRF)
  • Unigrams, bigrams and trigrams lemmatized
  • Verb POS, passive or not, presence
  • Verb class (10 classes) obtained by clustering verbs with frequency > 150
  • Grammatical triples from dependency tree

Results:

• Accuracy: SVM: 51.6% CRF: 50.4%
• Most relevant feature sets: bigrams, triples from dependency tree, verbs as well as structural features as history and section heading type (ngram 65,000 features vs 13,000 all other features)
• There is not always a direct correlation of annotator agreement and classifier performance: Experiment and Model have an higher F-score but low inter-annotator agreement
Scientific Discourse Characterization

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: Core Scientific Concepts, multi-class

Dealing with the case in which more than one Core Scientific Concept appears in a single sentence

**METHOD**
Bone marrow stromal cells were treated with AhR agonists and bacterial lipopolysaccharide (LPS) to mimic innate inflammatory cytokine responses.

**GOAL**

Multi-CoreSC Corpus

50 papers from the domain of cancer risk assessment
Environmental Health Perspectives (21), Carcinogenesis (15), Toxicological Sciences (9), Journal of Biological Chemistry (3), Occupational and Environmental Medicine (1), PlosOne (1)

• 8,501 sentences

Online at (SciXML format):
http://www.sapientaproject.com/wp-content/uploads/2016/05/consensus_annotated.zip
Scientific Discourse Characterization

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: Core Scientific Concepts, multi-class

Dealing with the case in which more than one Core Scientific Concept appears in a single sentence

**Multi-CoreSC CRA Corpus**
- 3 biology expert annotators
- weighted kappa > 0.55 for each ann. pair
- **12.5% of sentences obtained a multi-CoreSC label**
- **multi label conciliation procedure** to generate Gold Standard: lower number of labels across annotators in Gold Standard. Labels are ranked with respect to popularity and in case of equal popularity with respect to priority

**Is CoreSC CRF classifier domain indep.?**
Old: trained on ART corpus, **New**: trained and tested on CRA corpus

<table>
<thead>
<tr>
<th>Label</th>
<th>F-measure</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Old</td>
<td>New</td>
</tr>
<tr>
<td>Bac</td>
<td>52.6</td>
<td>76.7</td>
</tr>
<tr>
<td>Con</td>
<td>41.3</td>
<td>53.9</td>
</tr>
<tr>
<td>Exp</td>
<td>80.9</td>
<td>83.4</td>
</tr>
<tr>
<td>Goa</td>
<td>43.3</td>
<td>54.2</td>
</tr>
<tr>
<td>Hyp</td>
<td>11.6</td>
<td>13.9</td>
</tr>
<tr>
<td>Met</td>
<td>48.2</td>
<td>58.3</td>
</tr>
<tr>
<td>Mod</td>
<td>00.0</td>
<td>00.0</td>
</tr>
<tr>
<td>Mot</td>
<td>07.4</td>
<td>43.9</td>
</tr>
<tr>
<td>Obj</td>
<td>26.1</td>
<td>27.0</td>
</tr>
<tr>
<td>Obs</td>
<td>35.1</td>
<td>43.0</td>
</tr>
<tr>
<td>Res</td>
<td>47.5</td>
<td>68.3</td>
</tr>
</tbody>
</table>

Most influential features of CoreSC annotation are **domain specific**

**Object and Experiment**: only two categories that are consistently identified without domain adaptation
Scientific Discourse Characterization

Zone Analysis: Argumentative Zoning vs Core Scientific Concepts


AZ-II characterizes the ownership of the knowledge claims presented in the paper, thus identifying and motivating the new contributions of the author.

CoreSC describes the structure of the investigation characterizing the high level scientific concept presented in each part of the paper.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIM</td>
<td>Statement of specific research goal, or hypothesis of current paper</td>
<td>OWN_CONC</td>
<td>Findings, conclusions (non-measurable) of own work</td>
</tr>
<tr>
<td>NOV_ADV</td>
<td>Novelty or advantage of own approach</td>
<td>CO_DI</td>
<td>Comparison, contrast, difference to other solution (neutral)</td>
</tr>
<tr>
<td>CO_GRO</td>
<td>No knowledge claim is raised (or knowledge claim not significant for the paper)</td>
<td>OAP_WEAK</td>
<td>Lack of solution in field, problem with other solutions</td>
</tr>
<tr>
<td>OTHER</td>
<td>Knowledge claim (significant for paper) held by somebody else. Neutral description</td>
<td>ANTISUPP</td>
<td>Clash with somebody else’s results or theory, superiority of own work</td>
</tr>
<tr>
<td>PREV_OWN</td>
<td>Knowledge claim (significant) held by authors in a previous paper. Neutral description</td>
<td>SUPPORT</td>
<td>Other work supports current work or is supported by current work</td>
</tr>
<tr>
<td>OWN_MTHD</td>
<td>New knowledge claim, own work; methods</td>
<td>USE</td>
<td>Other work is used in own work</td>
</tr>
<tr>
<td>OWN_FAIL</td>
<td>A solution/method/experiment in the paper that did not work</td>
<td>FUT</td>
<td>Statements/suggestions about future work (own or general)</td>
</tr>
<tr>
<td>OWN_RES</td>
<td>Measurable/objective outcome of own work</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Corpora for the Conceptualisation and Zoning of Scientific Papers.
Schemata have complementary roles - it would be beneficial to annotate a text with respect to both schemata. In particular:

- **AZ-II** identifies knowledge claims that permeates several CoreSC concepts
- **CoreSC** has more granularity when dealing with content-related categories

Scientific Discourse Characterization

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: IMRAD


- Introduction > Methods > Results > Discussion
- Structure common to most health science journals
- Today more complex derived structures are often used


- first used in 1940s
- in 1970s 80% of compliant papers
- since 1980s most of health science papers are compliant
Scientific Discourse Characterization

Zone Analysis: Dr. Inventor Scientific Discourse Schema


Schema defined by annotating Computer Graphics papers, starting from AZ and CoreSC schemas (15 categories, then reduced to 5 top level + 2 second level)

- **Background**
- **Approach**
- **Challenge**
  - **Hypothesis**
  - **Goal**
- **Outcome**
- **Contribution**
- **Future Work**

Dr. Inventor Corpus

- 10 papers in **FLUID SIMULATION**
- 10 papers in **SKINNING**
- 10 papers in **CLOTH SIMULATION**
- 10 papers in **MOTION**

- 40 papers / 10,403 sentences
- Multi-layered annotations: *discursive structure*, citation purpose, summary sentence relevance
Scientific Discourse Characterization

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: Dr. Inventor Scientific Discourse Schema


Annotators

Distribution of sentence rhetorical class (over papers’ length)

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Background</th>
<th>Approach</th>
<th>Outcome</th>
<th>Future Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>4%</td>
<td>20%</td>
<td>58%</td>
<td>16%</td>
<td>2%</td>
</tr>
</tbody>
</table>

Annotation workflow

Annotation check after:
- 5 papers
- 15 papers
- 25 papers

Avg. annotator agreement K: 0.67

Online at: [http://sempub.taln.upf.edu/dricorpus](http://sempub.taln.upf.edu/dricorpus)
Scientific Discourse Characterization

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Dr. Inventor Scientific Discourse Schema


**CORPUS:** 8,777 sentences that have been manually associated to one of the 5 high level classes

**CLASSIFIERS:** Logistic regression, SVM (linear)

**FEATURES:** sentence position (only structural feat.), unigrams, bigrams, three-grams, dep. tree dept, num. and type of edges, dep. tree tokens, num and syntactic role of citations, category of previous sentence

**RESULTS:**
- in general the F-score of each category is proportional to the number of training instances
- Future Work has more strongly distinctive linguistic features than Challenge

<table>
<thead>
<tr>
<th>Category</th>
<th>Logistic Regression</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td>0.876</td>
<td>0.851</td>
</tr>
<tr>
<td>Background</td>
<td>0.778</td>
<td>0.735</td>
</tr>
<tr>
<td>Challenge</td>
<td>0.466</td>
<td>0.430</td>
</tr>
<tr>
<td>Future Work</td>
<td>0.675</td>
<td>0.496</td>
</tr>
<tr>
<td>Outcome</td>
<td>0.679</td>
<td>0.623</td>
</tr>
<tr>
<td><strong>Avg. F1:</strong></td>
<td><strong>0.801</strong></td>
<td><strong>0.764</strong></td>
</tr>
</tbody>
</table>

**Percentage of annotated sentences by category**
Scientific Discourse Characterization

Automated annotation of ABSTRACTS of scientific texts


**CORPUS:** 1,000 MedLine abstracts concerning Cancer Risk Assessment (7,985 sentences)
   On-line at: [http://www.cl.cam.ac.uk/~yg244/abstract_az.html](http://www.cl.cam.ac.uk/~yg244/abstract_az.html)

**3 ANNOTATION SCHEMAS:** (k measured over 1/3 of the corpus, three annotators)
   - Objective, Method, Results and Conclusion \((K=0.84) \rightarrow SVM \text{ acc: } 0.89\)
   - AZ, 7 categories \((K=0.85) \rightarrow SVM \text{ acc: } 0.90\)
   - CoreSC 11 categories \((K=0.50) \rightarrow SVM \text{ acc: } 0.81\)

**CLASSIFIERS:** Naïve Bayes, SVM with linear kernel (Weka)

**FEATURES:** location (10 equal parts), unigram, bigram, verb class (60 cluster of frequent verbs), grammatical triples from dependency tree, passive verb

**RESULTS:**
   - SVM outperforms Naïve Bayes in all cases (accuracy reported before)
   - Best features for all schemas: bigrams, verb and unigrams
   - Worse features for all schemas: history and voice worst (with abstract, the history of the categories is more varied and has less relevance than in the case in which we consider the whole text)
Automated annotation of ABSTRACTS of scientific texts


CORPUS: 51,000 MedLine abstracts with sentences divided in Objective, Method, Result and Conclusion

CLASSIFIERS: SVM (linear kernel), CRF

FEATURES: unigrams and bigrams also from next and previous sentence features, relative sentence location

RESULTS:
• CRF outperformed the SVM with features from previous and next sentence showing that is more adequate to classify sentences of scientific abstracts
• Since features are mainly based on lexical contents of annotated text (unigrams and bigrams), the accuracy strongly improves when a greater dataset is considered
Automated annotation of scientific texts: **ACTIVE LEARNING**

Guo, Y., Silins, I., Stenius, U., & Korhonen, A. (2013). *Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review*. Bioinformatics, 29(11)

**CORPUS**: 50 biomedical articles (8,171 sentences) annotated with AZ categories

**CLASSIFIER**: SVM (linear kernel)

Driven selection of new samples to consider to increase the training set by means of three strategies:

- **least confident sampling**: instance with more classification uncertainty
- **margin sampling**: instance with the smallest margin between the priors of the two most likely labelings
- **query-by-bagging**: a committee of models trained on subset of training instances is created and chosen the instance for which the committees disagree the most (most informative instance)

**FEATURES**: unigrams, bigrams, normalized section name, location inside section and paragraph, number of cits and table/figure references, verb class, tense, voice, dep. rels

**RESULTS**:  
- active learning with SVM trained on 6% of the corpus performs surprisingly well with the accuracy of 82%, just 2% lower than fully supervised learning

<table>
<thead>
<tr>
<th>Method</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random selection</td>
<td>0.73</td>
<td>0.76</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>Active learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Least confident</td>
<td>0.73</td>
<td>0.76</td>
<td>0.77</td>
<td>0.79</td>
<td>0.80</td>
<td>0.80</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>Margin</td>
<td>0.73</td>
<td>0.76</td>
<td>0.78</td>
<td>0.79</td>
<td>0.80</td>
<td>0.80</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>Query-by-bagging</td>
<td>0.73</td>
<td>0.77</td>
<td>0.78</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>0.80</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Fully supervised accuracy (8,171 instances): 0.84
Scientific discourse characterization

Automated annotation of scientific texts


- Use of Active learning and semi-supervised approaches to improve discursive sentence classification
- Active SVM outperforms the best supervised SVM with a statistically significant difference exploiting only a fraction of the training data


- Adding manually defined constraints to complement the statistical classification of sentences
- Two types of constraints are defined:
  - *lexical*: there is one or more reference to figures and tables, there is one or more citation, there are occurrences of specific word classes
  - *discursive*: is the first / last part of the paragraph or section
Overview of available datasets

**AZ Corpus**: 80 articles computational linguistics
http://www.cl.cam.ac.uk/~sht25/AZ_corpus.html

**ART Corpus**: 265 papers from the domains of chemistry and biochemistry

**MultiCoreSC CRA Corpus**: 50 papers from the domain of Cancer Risk Assessment
http://www.sapientaproject.com/wp-content/uploads/2016/05/consensus_annotated.zip

**Dr. Inventor Multi-layered Corpus**: 40 papers from the domain of Computer Graphics
http://sempub.taln.upf.edu/dricorpus

**MedLine Abstracts Corpus**: 1,000 MedLine abstracts on Cancer Risk Assessment
http://www.cl.cam.ac.uk/~yg244/abstract_az.html
Conclusions

• The characterization of scientific discourse provides valuable information to enhance several scientific text mining tasks like text quality assessment, information extraction, content retrieval and summarization.

• Zone Analysis is the most widespread approach to characterize scientific discourse, often at sentence level.

• Annotation schemas often offers complementary views by modeling different aspects of scientific discourse.

• Even if minimal, often annotation schemas need to be adapted to the specific domain of the scientific textual contents to characterize.

• Supervised approaches are widely explored: classifiers (Naïve Bayes, logistic regression, SVM) or sequence labeling approaches (CRF).

• A rich set of annotated corpora is freely available for further experimentation.
CITATION ANALYSIS

jays, robins and other birds”. These types of models have been used for hyponym discovery (Hearst, 1992; Roark and Charniak, 1998) meronym discovery (Berland and Charniak, 1999) and hierarchy building (Caraballo, 1999). These methods are very interesting but of limited applicability, because nouns that do not appear in known lexico-syntactic patterns cannot be learned.


In this paper we build upon the work of Riedel et al. (2013), which jointly learns continuous representations for knowledge base and textual relations. This common representation in the same
Outline

• Citations in scientific literature

• How citations are studied?
  – Citation network analysis
  – Citation function
  – Citation prediction and recommendation
  – Citation-based summarization

• Citation graphs

• Conclusions
Citations in scientific literature

Citations are the primary device used in scientific literature to relate a piece of work with other relevant (published) materials.

We cite papers to:

• **ground the arguments** and give the work factual basis
• **avoid plagiarism** (intellectual honesty)
• **attribute prior or unoriginal work and ideas** to the correct sources
• allow the reader to **determine independently whether the referenced material supports the author's argument** in the claimed way (demonstrate assessors and critics you have carried out the necessary research)
• enable the reader to independently evaluate the strength and validity of the material the author has used

**CITATION**

**Body of the paper**

- Strube and Ponzetto (2006) were the first to compute measures of semantic relatedness using Wikipedia. Their

**Bibliography**

Citation analysis

Citations in scientific literature

The elements of a citation

Each citation is a directed link from a **citing paper** to a **cited paper**

In this paper, we investigate the potential of weakly-supervised learning for Argumentative Zoning (AZ) of scientific abstracts. AZ is an approach to information structure which provides an analysis of the rhetorical progression of the scientific argument in a document (Teufel and Moens, 2002). It has been used to analyze scientific texts in various disciplines – including computational linguistics (Teufel and Moens, 2002), law, (Hachey and Grover, 2006), biology (Mizuta et al., 2006) and chemistry (Teufel et al., 2009) – and has proved useful for NLP tasks such as summarization (Teufel and Moens, 2002). Although the basic scheme is said to be discipline-independent (Teufel et al., 2009), its application to different domains has resulted in various modifications and laborious annotation exercises. This suggests that a weakly-supervised approach would be more practical than a fully supervised one for the real-world application of AZ.

The text of the **citing paper** surrounding an **in-line citation** and motivating the same citation is referred to as **citation context**

The excerpt of the **cited paper** that explains the actual contents cited by the **citing paper** surrounding is referred to as **cited span**
How citations are studied?

• Citation network analysis
• Citation function
• Citation prediction and recommendation
• Citation-based summarization
Citation analysis

Citation network analysis

Citation networks:
• *nodes*: papers
• *arcs*: directed from citing to cited paper

The more often a single paper is cited, the more important it seems to be
Bibliographic coupling network:
- **nodes**: papers
- **arks**: undirected, connect pairs of documents that share one or more cited documents

*Retrospective*: is limited to the papers cited by a pair of articles and cannot vary with time

The more often two papers are cited together, the more likely they are to be part of some research question or ongoing problem or conversation topic within the discipline

Co-citation network:
- **nodes**: papers
- **arks**: undirected, connect a pair of papers if they are cited by the same document(s)

*Non-retrospective*: may vary by new citations received by the papers in the future

Co-Citation Proximity Index (CPI) can be introduced to account for the placement of citations relative to each other. Documents co-cited at greater relative distances in the full text receive lower CPI values.

INTERACTIVE EXAMPLE: [http://jgoodwin.net/network/cites-slider.html](http://jgoodwin.net/network/cites-slider.html)
Citation analysis

Citation network analysis

Author coupling and co-citation networks

Author bibliographic coupling:
• nodes: authors
• arks: undirected, equals to the number of references that the publications of the pairs of authors have in common

Method to map the research activities of active authors themselves for a more realistic picture of the current state of research in a field

Author co-citation:
• nodes: authors
• arks: undirected, the number of times the pair of connected authors are cited together by the same article

Method to study the external and internal as well as recent and historical intellectual influences on the field
Citation analysis

Citation network analysis

How citation networks are exploited?

• identify “hot” areas and key authors (authors that are most collaborative or are most highly cited) → centrality, in-degree, out-degree

• community detection (meaningful communities of researchers) → clustering methods

• understand the research habits, trends, and topological patterns of the researchers

• spot and characterize productivity, patterns and trends

• provide complementary data to enhance the analysis of the contents of scientific publications
Citation network analysis

Microsoft Academic Graph

February 2016
- 47,742,000 papers with at least one reference or one citation (36.8% of total papers)
- 528,682,289 internal citations
- Each paper in the graph is cited on average 4.17 times

CiteSeer\textsuperscript{x} \textsubscript{10M}

October 2016
- 384,413 papers
- 1,751,463 internal citations
- Each paper in the graph is cited on average 4.56 times
Citation analysis

Citation network analysis

Improving detection of scientific topic evolution by citation network

(Scientific) Topic detection and evolution
Discover how and what topics change over time since the evolution of a topic in a specific time period can boost the investigation of other topics in subsequent periods

YEAR 1

TOPIC 1

TOPIC 2

YEAR 2

TOPIC 1

TOPIC 3

TOPIC 4

YEAR 3

TOPIC 1

TOPIC 3

TOPIC 5

YEAR 4

TOPIC 6

TOPIC 7

same topic
related topic
Once defined a number $k$ of topics, given a collection of documents, a topic detection method generates for each topic $z$ a vocabulary distribution so as to maximize the likelihood of the observed data.

Given $1 \geq a > b > 1/k$, a pair of topics $z(T)$ and $z(T-1)$ respectively computed over document collections at time $T$ and time $T-1$ is:

- **equal**: $p( z(T) | z(T-1) ) > a$
- **similar**: $b < p( z(T) | z(T-1) ) < a$
- **new**: $p( z(T) | z(T-1) ) < b$

$p( z(T) | z(T-1) )$ equal to $\text{sim}(z(T), z(T-1))$

Citation analysis

Citation network analysis

Improving detection of scientific topic evolution by citation network

Considering a collection of scientific paper spanning a number of years, in order to track year-by-year topic evolution, we can generate the topic of each year by different approaches:

- **Time independent topic evolution learning**
  \[
  Z(i-1) \rightarrow D(i-1) \rightarrow Z(i) \rightarrow D(i) \rightarrow Z(i+1) \rightarrow D(i+1)
  \]

- **Accumulative topic evolution learning**
  \[
  Z(i-1) \rightarrow D(i-1) \rightarrow Z(i) \rightarrow D(i) \rightarrow Z(i+1) \rightarrow D(i+1)
  \]

- **Citation-aware topic evolution learning**
  \[
  Z(1) \rightarrow L_{D(2)} \rightarrow D(2) \rightarrow L_{D(i)} \rightarrow D(i)
  \]

Citation network analysis

Improving detection of scientific topic evolution by citation network

Citation-aware topic evolution learning

\[ Z(t) = \arg \max_{Z(t)} \prod_{d \in D(t) \cup L_D(t)} p'(d|Z(t)) , \]

where

\[ p'(d|Z(t)) = \lambda \cdot p(d|Z(t)) + (1 - \lambda) \cdot \sum_{d_i \in L_d} \gamma_{d_j} \cdot p'(d_j|Z(t)) \]

• not all citations are equally important (only few can be related to the topic of the citing paper)
• when historical papers are cited, some out-of-date topic may be wrongly considered

Inheritance topic model

The autonomous part and the inherited part of a paper (cited papers) are learned independently

Citation network analysis

Improving detection of scientific topic evolution by citation network

Evaluation: 650,918 computer and information science papers from CiteSeer from 1993 to 2008

Evolution of 30 topics studied with different approaches

<table>
<thead>
<tr>
<th>Time independent topic evolution learning</th>
<th>Accumulative topic evolution learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation unaware</td>
<td>Citation aware</td>
</tr>
</tbody>
</table>

- **Citation-unaware topic evolution learning**
  - Historical topics tend to dominate; difficult to detect new ones
  - Produce less new topics with respect to ITM since old, cited papers are treated as new ones

- **Citation-aware topic evolution learning**
  - Large number of new topics. Several noisy topics
  - Reach a good balance between new and old topics enabling detection of new ones

- **Inheritance topic model**
  - Produce less new topics with respect to ITM since old, cited papers are treated as new ones
  - Historical topics tend to dominate; difficult to detect new ones

---

Citation network analysis

Co-authorship network

reflects the personal link between scientists

- maths: 70,975 authors and 70,901 papers
- neuroscience: 209,293 authors, 210,750 papers

Results:
- degree distribution: power-law, is a scale free network
- the average node separation slightly decreases over time: more internal inks are produced with time (co-authorships) increasing network interconnectivity and decreasing diameter
- the average degree increases with time
- node selection is governed by preferential attachment

Citation analysis

Citation network analysis

Author co-citation network

- 15 journals classified in ISI’s Web of Science dealing with Philosophy and History
- 12,510 articles dating from 1956 with over 300,000 citations between them

Authors co-citation graph shows intellectual influences of individual authors, clustering them by discipline

Source: http://www.scottbot.net/HIAL/wp-content/uploads/2013/06/FullPageMapOfHPS.png
Citation analysis

Citation function

Many research impact and quality indexes are based on citation counts but...

Not all citations are equal!

\[
\text{Impact Factor (corrected)} = \frac{\# \text{ times your work is cited} - \# \text{ citations that actually trash your work} - \# \text{ times you cited yourself} (\text{nice try}) - \# \text{ times you were cited just to pad the introduction section}}{\# \text{ original articles you've written} + \# \text{ articles you were included in out of pity or politics} + \# \text{ not-so-original articles you've written copied and pasted}}
\]

There are different motivation that could explain why an author cites other pieces of research

- Judge: criticize a work express contrary or negative judgments
- Explore: investigate used as a starting point for the work described
- Show results: highlight a positive result

In this paper we build upon the work of Riedel et al. (2013), which jointly learns continuous representations for knowledge base and textual relations. This common representation in the same investigations used as a starting point for the work described.
Many research impact and quality indexes are based on citation counts but...

**Not all citations are equal!**

The approach presented by [1] presents several limitations.

We parse text by means of the concept extraction system presented in [1].

We compare our system with the concept extraction performance of [1].

Our systems extracts concepts by implementing the approach presented by [1].

The system presented by [1] can’t be easily adapted to texts from different domains.

The main drawback of [1] is its impossibility to scale to large systems.

This paper has 6 citations!

Half of the citations of this paper **criticize** aspect of the work presented.

Two citations of this paper **use** the approach / tool presented.

One citation of this paper **compares** the approach / tool presented.
Many research impact and quality indexes are based on citation counts but...

Not all citations are equal!

In order to understand why a paper is cited it is fundamental to correctly identify the citation context, that is the text excerpt(s) of the citing paper that explains and motivates the citation.

The citation context:

- may include sentences surrounding the one where the in-line citation occurs
- only part of the sentence where the in-line citation occurs can contribute to motivate the citation.
Several annotation schemas have been proposed to characterize the function of citations.


**4 dimensions for citation characterization**

<table>
<thead>
<tr>
<th>Conceptual</th>
<th>If a concept or a theory of the cited paper is used directly or indirectly in the citing paper in order to lay foundations to build on it or to contribute to the citing paper, then the citation is a conceptual one.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational</td>
<td>When a concept or theory id referred to as tool... [or] when it borrows mathematical or physical techniques, results, references, or conclusions from the cited paper.</td>
</tr>
<tr>
<td>Organic</td>
<td>Those [papers] from which concepts or theories are taken to lay the foundations of the citing paper, or papers from which certain results (including numerical ones) are taken to develop the ideas in the citing paper, or papers which help to better understand certain concepts in the citing paper.</td>
</tr>
<tr>
<td>Perfunctionary</td>
<td>Those [papers] which describe alternative approaches are not utilized in the citing papers.... references which are used to indicate the fact that a certain method employed is routine in the literature, and references which merely contribute to the chronological context of the citing paper.</td>
</tr>
<tr>
<td>Evolutionary</td>
<td>[The paper] provides a concept or theory to build on, or a mathematical technique to use, or results of an analysis which is used in the development of the citing paper, or notation used in the citing paper.</td>
</tr>
<tr>
<td>Juxtapositional</td>
<td>[The paper] refers to alternative approaches... [and] refers to other analysis used in the citing paper only to make comparisons, refers to other works which may help to clarify some ideas but do not contribute to the development of the citing paper, or refers to a paper only for references given in the latter.</td>
</tr>
<tr>
<td>Confirmative</td>
<td>A reference is confirmative if the author of the citing paper considers the paper referred to as correct.</td>
</tr>
<tr>
<td>Negative</td>
<td>The author of the citing paper is not certain about the correctness of the cited paper.</td>
</tr>
</tbody>
</table>

**Use of theory**
- use of technical method
- work is crucially needed for understanding of citing article
- just a general acknowledgement

**Own work**
- own work is an alternative to cited work
- own work is based on the cited work
- the work confirm the cited paper

**Negative**
- cited paper are criticized

**40% of citations are Perfunctionary**
(30 Articles in Physical Review, Published on Theoretical High Energy Physics from 1968 to 1972)
Citation function

Several annotation schemas have been proposed to characterize the function of citations


13 classes for citation characterization

1. Cited source substantiates a statement or assumption, or points to further information.
2. Cited source is mentioned in the introduction or discussion as part of the history and state of the art of the research question under investigation.
3. Cited source contains the data (pertaining to the discipline of the citing article) which are used for comparative purposes in tables and statistics.
4. Cited source contains the data pertaining to the discipline of the citing article) which are used sporadically in the citing text.
5. Cited source is positively evaluated.
6. Cited source contains the method used.
7. Cited source contains the concepts, definitions, interpretations used (and pertaining to the discipline of the citing article).
8. Cited source is the specific point of departure for the research question investigated.
9. Results of citing article disprove, put into question the data as interpretation of cited source.
10. Cited source is negatively evaluated.
11. Results of citing article prove, verify, substantiate the data or interpretation of cited source.
12. Results of citing article furnish a new interpretation, explanation of the data of the cited source.
13. Cited source contains data and material (from other disciplines than citing article) which is used sporadically in the citing text, in tables or statistics.

80% of citations belong to the first category: Cited source substantiates a statement or assumption, or points to further information.

0.8% of citations criticize the cited paper.

Citation function

Several annotation schemas have been proposed to characterize the function of citations.


4 top level categories for citation characterization

<table>
<thead>
<tr>
<th>Weakness</th>
<th>Authors point out a weakness in cited work.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast</td>
<td>Authors make contrast/comparison with cited work (4 categories)</td>
</tr>
<tr>
<td>CoCoGM</td>
<td>Contrast/Comparison in Goals or Methods (neutral)</td>
</tr>
<tr>
<td>CoCoR0</td>
<td>Contrast/Comparison in Results (neutral)</td>
</tr>
<tr>
<td>CoCo</td>
<td>Unfavourable Contrast/Comparison (current work is better than cited work)</td>
</tr>
<tr>
<td>CoCoXY</td>
<td>Contrast between 2 cited methods</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Positive</th>
<th>Authors agree with/make use of/show compatibility or similarity with cited work (6 categories),</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBas</td>
<td>author uses cited work as starting point</td>
</tr>
<tr>
<td>PUse</td>
<td>author uses tools/algorithms/data</td>
</tr>
<tr>
<td>PModi</td>
<td>author adapts or modifies tools/algorithms/data</td>
</tr>
<tr>
<td>PMot</td>
<td>this citation is positive about approach or problem addressed (used to motivate work in current paper)</td>
</tr>
<tr>
<td>PSim</td>
<td>author's work and cited work are similar</td>
</tr>
<tr>
<td>PSup</td>
<td>author's work and cited work are compatible/provide support for each other</td>
</tr>
</tbody>
</table>

| Neutral           | Function of citation is either neutral, or weakly signalled, or different from the three functions stated above |

Corpus CitRAZ: 26 conference articles – 584 citations from Computation and Language archive

<table>
<thead>
<tr>
<th>Neut</th>
<th>PUse</th>
<th>CoCoGM</th>
<th>PSim</th>
<th>Weak</th>
<th>CoCoXY</th>
<th>PMot</th>
<th>PModi</th>
<th>PBas</th>
<th>PSup</th>
<th>CoCoGM0</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.7%</td>
<td>15.8%</td>
<td>3.9%</td>
<td>3.8%</td>
<td>3.1%</td>
<td>2.9%</td>
<td>2.2%</td>
<td>1.6%</td>
<td>1.5%</td>
<td>1.1%</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

Online at: [http://www.cl.cam.ac.uk/~sht25/Project_Index/Citraz_Index.html](http://www.cl.cam.ac.uk/~sht25/Project_Index/Citraz_Index.html) [http://www.cl.cam.ac.uk/~sht25/CFC.html](http://www.cl.cam.ac.uk/~sht25/CFC.html)
Several annotation schemas have been proposed to characterize the function of citations.


**Classifiers:** K-nearest neighbours classifier

**Citation features:**
- grammar (POS-based) with 1762 cue-phrases (Teufel, 1999)
- POS-based recognizer for *agents* and recognizer for *actions* that these agents perform (Teufel, 1999)
- **892 cue-phrases** (about 75 per citation function, identified by annotators)
- **verb tense and voice**
- **modality** (whether or not a main verb is modified by an auxiliary, and which auxiliary it is)
- **location of sentence** in the whole paper and in the section or paragraph
- **self citations**

**Results**

4 top classes

<table>
<thead>
<tr>
<th></th>
<th>Weakness</th>
<th>Positive</th>
<th>Contrast</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>.80</td>
<td>.75</td>
<td>.77</td>
<td>.81</td>
</tr>
<tr>
<td>R</td>
<td>.49</td>
<td>.65</td>
<td>.52</td>
<td>.90</td>
</tr>
<tr>
<td>F</td>
<td>.61</td>
<td>.70</td>
<td>.62</td>
<td>.86</td>
</tr>
</tbody>
</table>

All classes

<table>
<thead>
<tr>
<th></th>
<th>Weak</th>
<th>CoCoGM</th>
<th>CoCoR0</th>
<th>CoCo-</th>
<th>CoCoXY</th>
<th>PBas</th>
<th>PUse</th>
<th>PModi</th>
<th>PMot</th>
<th>PSim</th>
<th>PSup</th>
<th>Neut</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>.78</td>
<td>.81</td>
<td>.77</td>
<td>.56</td>
<td>.72</td>
<td>.76</td>
<td>.66</td>
<td>.60</td>
<td>.75</td>
<td>.68</td>
<td>.83</td>
<td>.80</td>
</tr>
<tr>
<td>R</td>
<td>.49</td>
<td>.52</td>
<td>.46</td>
<td>.19</td>
<td>.54</td>
<td>.46</td>
<td>.61</td>
<td>.27</td>
<td>.64</td>
<td>.38</td>
<td>.32</td>
<td>.92</td>
</tr>
<tr>
<td>F</td>
<td>.60</td>
<td>.64</td>
<td>.57</td>
<td>.28</td>
<td>.62</td>
<td>.58</td>
<td>.63</td>
<td>.37</td>
<td>.69</td>
<td>.48</td>
<td>.47</td>
<td>.86</td>
</tr>
</tbody>
</table>
Citation analysis

Citation function

Several annotation schemas have been proposed to characterize the function of citations


6 classes for citation characterization

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criticism</td>
<td>Criticism can be positive or negative. A citing sentence is classified as “Criticizing” when it mentions the weakness/strengths of the cited approach, negatively/positively criticizes the cited approach, negatively/positively evaluates the cited source.</td>
</tr>
<tr>
<td>Comparison</td>
<td>A citing sentence is classified as “Comparison” when it compares or contrasts the work in the cited paper to the author’s work. It overlaps with the first category when the citing sentence says one approach is not as good as the other approach. In this case we use the first category.</td>
</tr>
<tr>
<td>Use</td>
<td>A citing sentence is classified as “Use” when the citing paper uses the method, idea or tool of the cited paper.</td>
</tr>
<tr>
<td>Substantiation</td>
<td>A citing sentence is classified as “Substantiating” when the results, claims of the citing work substantiate, verify the cited paper and support each other.</td>
</tr>
<tr>
<td>Basis</td>
<td>A citing sentence is classified as “Basis” when the author uses the cited work as starting point or motivation and extends on the cited work.</td>
</tr>
<tr>
<td>Neutral</td>
<td>A citing sentence is classified as “Neutral” when it is a neutral description of the cited work or if it doesn’t come under any of the above categories.</td>
</tr>
</tbody>
</table>

Corpus: 3,271 citations from ACL Anthology Network Corpus, annotated with respect to polarity and purpose

Online at: http://clair.si.umich.edu/corpora/citation_sentiment_umich.tar.gz
Several annotation schemas have been proposed to characterize the function of citations.


Classifiers: CRF

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demonstrative determiners</td>
<td>Takes a value of 1 if the current sentence contains a <em>demonstrative determiner</em> (this, these, etc.), and 0 otherwise.</td>
</tr>
<tr>
<td>Conjunctive adverbs</td>
<td>Takes a value of 1 if the current sentence starts with a <em>conjunctive adverb</em> (However, Furthermore, Accordingly, etc.), and 0 otherwise.</td>
</tr>
<tr>
<td>Position</td>
<td>Position of the current sentence with respect to the citing sentence. This feature takes one of four values: -1, 0, 1, and 2.</td>
</tr>
<tr>
<td>Contains Closest Noun Phrase</td>
<td>Takes a value of 1 if the current sentence contains closest noun phrase (if any) immediately before the reference position in the citing sentence, and 0 otherwise. This noun phrase often is the name of a method, a tool, or corpus originating from the cited reference.</td>
</tr>
<tr>
<td>2-3 grams</td>
<td>The first bigram and trigram in the sentence (<em>This approach, One problem with</em>, etc.).</td>
</tr>
<tr>
<td>Contains Other references</td>
<td>Takes a value of 1 if the current sentence contains references other than the target, and 0 otherwise.</td>
</tr>
<tr>
<td>Contains a Mention of target reference</td>
<td>Takes a value of 1 if the current sentence contains a mention (explicit or anaphoric) of the target reference, and 0 otherwise.</td>
</tr>
<tr>
<td>Multiple references</td>
<td>Takes a value of 1 if the citing sentence contains multiple references, and 0 otherwise. If the citing sentence contains multiple references, it becomes less likely that the surrounding sentences are related.</td>
</tr>
</tbody>
</table>

Results:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRFs</td>
<td>98.5%</td>
<td>82.0%</td>
<td>89.5%</td>
</tr>
<tr>
<td>ALL</td>
<td>30.7%</td>
<td>100.0%</td>
<td>46.9%</td>
</tr>
<tr>
<td>CS-ONLY</td>
<td>88.0%</td>
<td>74.0%</td>
<td>80.4%</td>
</tr>
<tr>
<td>SVM</td>
<td>92.0%</td>
<td>76.4%</td>
<td>83.5%</td>
</tr>
</tbody>
</table>

Lexical features are more important than structural features.
Citation function

Several annotation schemas have been proposed to characterize the function of citations.


**Classifiers:** SVM with linear kernel

**Citation features:** (ordered by relevance)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference count</td>
<td>The number of references that appear in the citation context.</td>
</tr>
<tr>
<td>Is Separate</td>
<td>Whether the target reference appears within a group of references or separate (i.e. single reference).</td>
</tr>
<tr>
<td>Closest Verb / Adjective / Adverb</td>
<td>The lemmatized form of the closest verb/adjective/adverb to the target reference or its representative or any mention of it. Distance is measure based on the shortest path in the dependency tree.</td>
</tr>
<tr>
<td>Self Citation</td>
<td>Whether the citation from the source paper to the target reference is a self citation.</td>
</tr>
<tr>
<td>Contains 1st/3rd PP</td>
<td>Whether the citation context contains a first/third person pronoun.</td>
</tr>
<tr>
<td>Negation</td>
<td>Whether the citation context contains a negation cue. The list of negation cues is taken from the training data of the <em>SEM 2012 negation detection shared task</em> (Morante and Blanco, 2012).</td>
</tr>
<tr>
<td>Speculation</td>
<td>Whether the citation context contains a speculation cue. The list is taken from Quirk et al. (1985)</td>
</tr>
<tr>
<td>Closest Subjectivity Cue</td>
<td>The closest subjectivity cue to the target reference or its representative or any anaphoric mention of it. The list of cues is taken from OpinionFinder (Wilson et al., 2005)</td>
</tr>
<tr>
<td>Contrary Expressions</td>
<td>Whether the citation context contains a contrary expression. The list is taken from Biber (1988)</td>
</tr>
<tr>
<td>Section</td>
<td>The headline of the section in which the citation appears. We identify five title categorizes: 1) Introduction, Motivation, etc. 2) Background, Prior Work, Previous Work, etc. 3) Experiments, Data, Results, Evaluation, etc. 4) Discussion, Conclusion, Future work, etc.. 5) All other section headlines. Headlines are identified using regular expressions.</td>
</tr>
<tr>
<td>Dependency Relations</td>
<td>All the dependency relations that appear in the citation context. For example, <em>nsubj(outperform, algorithm)</em> is one of the relations extracted from &quot;This algorithm outperforms the one proposed by...&quot;. The arguments of the dependency relation are replaced by their lemmatized forms. This type of features has been shown to give good results in similar tasks (Athar and Teufel, 2012a).</td>
</tr>
</tbody>
</table>

**Results:**

- Structural features and features characterizing the words surrounding the citation to classify are the most important.
- Considering the citation context improves classification of subjective categories (exp. Negative)
Several annotation schemas have been proposed to characterize the function of citations.


**Corpus:** 8,736 citations from 310 research papers taken from the ACL Anthology, tagged manually as positive, negative or objective

Online at: http://cl.awaisathar.com/citation-sentiment-corpus/

**Features:**
- unigrams, bigrams and trigrams adding to the lemma also the POS of every token
- name of the primary author of the cited paper
- science lexicon: 83 polar phrases which have been manually extracted from the development set of 736 citations
- presence of subjectivity clues
- number of adjectives, adverbs, pronouns, modals and cardinals
- number of negation phrases and valence shifter (Opinion Finder)
- dependency tree triples

Dependency tree used to identify the clause of the sentence where the citation occurs

Negated (suffix ‘_neg’) the two lemmas after a negation expression (Opinion Finder)
Citation analysis

Citation function

Several annotation schemas have been proposed to characterize the function of citations


**Corpus**: 8,736 citations from 310 research papers taken from the ACL Anthology, tagged manually as positive, negative or objective


**Algorithm**: SVM

**Result**:

- n-grams and dependency relations are sufficient to model lexical structure that can characterize the polarity of citations
- scientific lexicon, word level features, sentence splitting and negation does not help


**Corpus**: 852 papers which cite the top 20 target papers. Citation context sentences are identified and marked as negative, positive, objective/neutral.

Several annotation schemas have been proposed to characterize the function of citations.


**23 properties for citation characterization**

<table>
<thead>
<tr>
<th>Factual relationships</th>
<th>Rhetorical relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>cito:cites</td>
<td>cito:confirms</td>
</tr>
<tr>
<td>cito:citesAsAuthority</td>
<td>cito:credits</td>
</tr>
<tr>
<td>cito:citesAsMetadataDocument</td>
<td>cito:extends</td>
</tr>
<tr>
<td>cito:citesAsSourceDocument</td>
<td>cito:obtainsSupportFrom</td>
</tr>
<tr>
<td>cito:citesForInformation</td>
<td>cito:supports</td>
</tr>
<tr>
<td>cito:isCitedBy</td>
<td>cito:updates</td>
</tr>
<tr>
<td>cito:obtainsBackgroundFrom</td>
<td></td>
</tr>
<tr>
<td>cito:sharesAuthorsWith</td>
<td></td>
</tr>
<tr>
<td>cito:usesDataFrom</td>
<td></td>
</tr>
<tr>
<td>cito:usesMethodIn</td>
<td></td>
</tr>
</tbody>
</table>

Part of the *Semantic Publishing and Referencing Ontologies*, includes 41 properties for citation characterization in its most recent version (03/07/2015)
Several annotation schemas have been proposed to characterize the function of citations.

**CRITICISM**
- WEAKNESS: A weakness in a cited work may refer to some restriction, its inappropriateness in the case considered, a requirement, its difficulty, its computational cost, etc.
- STRENGTH: A strength in a cited work may refer to its easiness of use, its little computational cost, its speed, its novelty, etc.
- EVALUATION: Some citations do not only state a strength or a weakness of the cited paper, but provide the author’s evaluation of the research, by opposing a strength with a weakness or by giving his opinion in an explicit (or subtle) way.
- OTHER: If a citation can be considered a CRITICISM, but cannot be included in the previous sub-purposes, then it should be annotated as CRITICISM_OTHER.

**COMPARISON**
- SIMILARITY: The comparison focuses on the similarities with the author’s work.
- DIFFERENCE: The comparison focuses on the differences with the author’s work.

**USE**
- METHOD: If the author uses the method, technique, or algorithm developed by the cited paper.
- DATA: If the author uses the data produced by the cited paper.
- TOOL: If the author uses a tool or software package developed by the cited paper.

**SUBSTANTIATION**
- A citing sentence is classified as SUBSTANTIATION when the cited paper and the citing paper support each other.

**BASIS**
- PREVIOUS OWN: The author bases the current research on his own previous work.
- OTHERS’ WORK: The author bases the current research on others’ previous work.
- FUTURE WORK: Future work can be developed based on the cited work.

**NEUTRAL**
- DESCRIPTION: If the citation is a neutral description of the cited work.
- REFERENCE FOR MORE INFO: If the author refers to a work for obtaining more detailed information about a particular subject.
- COMMON PRACTICE: When other author’s work are cited as common practices in the knowledge field.
- OTHER: Other reasons for neutral citations.

---

**Dr. Inventor Corpus**
- 40 Computer Graphics articles
- 1,575 citations

Online at: [http://sempub.taln.upf.edu/dricorpus](http://sempub.taln.upf.edu/dricorpus)
Several annotation schemas have been proposed to characterize the function of citations.


### Coarse and fine-grained labels for citation type

<table>
<thead>
<tr>
<th>Citation Type</th>
<th>Fine-grained Label</th>
<th>Coarse Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related work</td>
<td>0</td>
<td>Incidental</td>
</tr>
<tr>
<td>Comparison</td>
<td>1</td>
<td>Incidental</td>
</tr>
<tr>
<td>Using the work</td>
<td>2</td>
<td>Important</td>
</tr>
<tr>
<td>Extending the work</td>
<td>3</td>
<td>Important</td>
</tr>
</tbody>
</table>

**Corpus:** 465 citations from ACL anthology

Considering direct citations:

Online baselines include Top-1 Perceptron (Collins, 2002), Top-1 Passive-Aggressive (PA), and k-best PA (Crammer & Singer, 2003; McDonald et al., 2004).

and indirect citations:

We implemented the MXPOST tagger and integrated it with our algorithm.

### Features:

- direct citations (total and per section)
- indirect citations (total and per section)
- author overlap
- is considered helpful
- in table or figure caption (we’re comparing)
- number of direct citations over all the direct citations
- tf-idf similarity between abstracts (citing / cited)
- page rank
- number of citing papers after transitive closure
- research field of paper

**Evaluation:** SVM accuracy: 0,93

**Algorithms:** SVM (RBF k.) and random forest

Most informative feature: direct citations (total and per section)
Followed by: author overlap, is considered helpful, number of direct citations over all the direct citations, research field of paper
Given an in-line citation placeholder, predict (recommend) which is the paper that should be cited.

Citation analysis

It analyses the content and conceptual structure of scientific articles with an ontology-based annotation scheme, the Core Scientific Concepts scheme (CoreSC). Closely related to this approach is the multidimensional scheme of Nawaz (2010), tailored to bioevents, and the works of De Waard (2009) in classifying sentences in 5 epistemic types and White (2011), who concentrates on identifying hypothesis, explanations and evidence in the biomedical domain.

Search query

It analyses the content and conceptual structure of scientific articles with an ontology-based annotation scheme the Core Scientific Concepts scheme (CoreSC).

4/12/2016

Cited paper not retrieved among the first 10 results.

No results retrieved.

Cited paper not retrieved among the first 10 results.
Several facets can contribute to identify the best cited paper match:

- features **local to the citation context** (e.g. papers with similar citation contexts)
- features **global of the whole document** (e.g. papers with similar title, abstract, shared keywords or authors)
- **user preferences** (i.e. publication and citation history of the author, user profile in a bibliography management system)
- **citation network** (i.e. paper-citation matrix)
Citation prediction and recommendation

Given an in-line citation placeholder, predict (recommend) which is the paper that should be cited

Dealing with citation prediction / recommendation...

• **Huge search space**: progressive reduction of candidate set (lightweight cited papers selection methods for a first coarse-grained selection, candidate cited paper clustering)

• **Neural models**: estimate the probability that, given a word from the citation context, a document is cited by jointly learning neural representations (embeddings) of words from citation contexts and cited documents

• **Citation context identification**: models to identify in a paper the candidate citation contexts and, for each of them, the list of top-n candidate cited papers

• **Citation motivation**: explain why a certain paper should be cited in a given citation context

Online tool that implements distinct approaches of document level and context level citation recommendation - [http://refseer.ist.psu.edu/](http://refseer.ist.psu.edu/)
The contexts of the citations of a specific paper provide useful information concerning the **core topic of the paper** together with **opinions of the research community** on the piece of work.

"Since citing sentences appear to be somewhat more focused than the abstract and contain additional information not in the abstract, they could be useful as a supplement"  
*Elkiss, A. et al. 2008). Blind men and elephants: What do citation summaries tell us about a research article?*

"The inclusion of citation-related information brings to the generation of better summaries."  
*Ronzano, F. et a. An Empirical Assessment of Citation Information in Scientific Summarization.*

**Citation-based summarization**

Past studies already presented neural-network approaches to identify named entities [1].

We parse text by means of the concept extraction system presented in [1].

[1] propose new approach to spot Named Entities in legal texts.

The approach described in [1] shows several limitations with respect to scalability.

"More details and examples of how summarization systems exploit citation-related information will be presented in the following part of this tutorial dealing with summarization of scientific literature."
Citation analysis

Citation graphs

Microsoft Academic Graph

528,682,289 internal citations, each paper in the graph is cited on average 4.17 times
https://academicgraphwe.blob.core.windows.net/graph-2016-02-05/index.html

High-energy physics citation network (2003 KDD cup)
Arxiv HEP-PH (high energy physics phenomenology) citation graph is from the e-print arXiv
January 1993 → April 2003 (124 months)
https://academicgraphwe.blob.core.windows.net/graph-2016-02-05/index.html

Patent citation network (2005 KDD cup)
16,522,438 citations, all citations made by patents granted between 1975 and 1999
January 1, 1963 → December 30, 1999 (37 years)
https://academicgraphwe.blob.core.windows.net/graph-2016-02-05/index.html

CiteSeer citation network
1,017,457 papers with 10,760,318 citations (Oct. 2013)

ACL Anthology Network
21,212 papers with 110,976 citations (Dec. 2013)

OpenCitations
1,740,050 bibliographic resources with 2,201,568 citations (Dec. 2016)
http://opencitations.net/ (RDF dataset / SPAR ontologies / main crawled source: Europe PMC)
Conclusions

• Citations represent a primary device of scientific literature useful to issue explicit author-created links among publications

• Both the network of citations and the textual contents of citation contexts are exploited in many different tasks including: research collaboration analysis, topic analysis and evolution, citation recommendation, scientific document summarization

• Besides citation counts, the (complex task of) characterization of the purpose of citations can provide deeper insights on the quality of scientific publications and the feedback of the research community

• Citation recommendation system can complement pure scientific literature search engines in helping to cope with scientific information overload

• A rich collection of citation datasets, including citation networks and corpora of citations annotated with respect to their sentiment and purpose, is freely available for further experimentation
SCIENTIFIC DOCUMENT SUMMARIZATION
Outline

• Document summarization overview

• Summarizing scientific articles
  – Information extraction and template-based generation
  – Indicative-informative summaries
  – Fact-based citation summaries (C-LexRank)
  – Impact-driven summaries
  – Non-explicit citations in summaries
  – Improving summary coherence
  – Generating state-of-the-art reports

• Summarizing patents

• Conclusions
What is a summary?
A presentation of the substance of a body of material in a condensed form or by reducing it to its main points; an abstract.
A short text containing the essential information of a document.

What is a summarizer?
An algorithm that selects and presents the most important content of a document.
Summaries should take into account a number of input factors such as the audience/reader of the summary.

**Types of Summaries**

- **one / several documents**
  - single document summarization
  - multi-document summarization

- **summary purpose**
  - indicative summary (what is text about?)
  - informative summary (gives information from text)

- **content**
  - generic summary (main points)
  - user-focused (answers a query)

- **input language**
  - monolingual
  - cross-lingual

- **formulation**
  - extract (sentences from input documents)
  - abstract ((quasi) new text )
Scientific document summarization

Document summarization overview

Summarization by sentence extraction
Extract from the input document the subset of sentences that contain the most important information

• General method to produce “extracts” of size N
  a) \( S = \{\} \)
  b) Associate to each sentence a score and put them in list L
  c) Sort sentences in L by score (in ascending order)
  d) While size of \( S < N \), put next sentence in L in S
  e) Show sentences in S in the order they appear in the original text

• Compression parameter
  – size in number of words of the summary
  – compression rate: % of the words or sentences
Word-distribution measures (Luhn'1958, Nenkova and Vanderwende, 2005)
  – Term/Word frequency

Document structure (Edmundson, 1969; Lin and Hovy, 1998)
  – Position of sentence in document
  – Relation of sentence to title, abstract, keywords, etc.

Presence of specific vocabulary (Paice, 1990)
  – Formulaic-expressions, key-words, etc.

Centrality information (Barzilay and Elhadad, 1997; Radev et al, 2000; Saggion and Gaizauskas, 2004)
  – Word-based sentence-sentence relations, centroid
  – Co-reference

Rhetorical information (Marcu, 1998; Ono et al., 1994)
  – How argument develops in sentences (more / less central)

Semantic (Saggion and Lapalme, 2002; Jones and Paice, 1993)
  – Domain/Topic template / Information types to cover in summary

External (Tombros et al, 1998)
  – Query / User Knowledge
Document summarization overview

Summarization by sentence extraction: superficial techniques

Information about sentence relevance can be provided by: word/term repetition and document structure

Word/term repetition

Over pre-processed corpora (stemming, stop-words removal), the inverse document frequency can be used to assess word relevance

\[
relevance(t) = tf(t) \times idf(t) \\
idf(term) = \log(\frac{NUMDOC}{NUMDOC(term)})
\]

Document structure

- Position of sentence in document
  - in news give relevance to lead-paragraph
  - in scientific discourse give relevance to sentences under specific section headings
  - learn optimal positions in a given textual genre

- Title / Query sentence relevance
  - similarity between sentence and document title or user need expressed in a query (cosine, jaccard, etc.)
  - Information retrieval techniques are useful here

(Edmundson, 1969; Lin and Hovy, 1998)
(Tombros et al, 1998)
Summarization by sentence extraction: superficial techniques

More or less fixed vocabulary indicates the presence of important information in text

Cue-phrases, indicative expressions, formulaic expressions, etc.

- dictionary with expressions (literal or patterns)
  - *in this {paper / work / article} we....*
  - *{our / my} {results / findings /...} demonstrate ....*
- may be organized in categories (results, conclusion, etc.)
- expressions may be weighted
- expressions might be learnt from corpora

(Paice, 1990; Jones and Paice, 1993)
No single source of information will produce the best scoring schema: usually features have to be combined

- Features can be combined "a la Edmundson" to score sentences
  
  \[ \text{Weight}(S) = \alpha \cdot \text{Title}(S) + \beta \cdot \text{Cue}(S) + \gamma \cdot \text{Keyword}(S) + \delta \cdot \text{Position}(S) \]

- Given training data
  - A scoring function can be learnt (regression)
  - A sentence classification function (extract | non-extract) can be learnt

Document summarization overview

Summarization by sentence extraction: graph-based techniques

Lexical similarity between sentences in the document tell us about their relevance: text is represented as a connected structure (unlike other superficial approaches)

- Text represented as a graph
  - vertices are “meaning” units such as words or sentences
  - edges are connections between units

- Inspired by the PageRank algorithm (Page et al. 1998) several summarization algorithms were proposed
  - LexRank (Erkan & Radev, 2004)
  - TextRank (Mihalcea & Tarau, 2004)

(Erkan and Radev, 2004; Mihalcea & Tarau, 2004)
Document summarization overview

Summarization by sentence extraction: graph-based techniques

**Page Rank**

conceived to rank Web pages by relevance

- Web pages form a directed graph
- PageRank computes the relevance (PageRank score) of each Web page thanks to the recursive analysis of the connectivity of the complete network:

\[
PR(i) = \frac{1-d}{N} + d \sum_{j=1}^{N} \frac{PR(j)}{C(j)}
\]

- PageRank values of page \(j\) connected to \(i\)
- Number of links going out of page \(j\)
- Damping factor
Scientific document summarization

Document summarization overview

Summarization by sentence extraction:
graph-based techniques at document level

Text graphs for summarization seek to associate a weight
to sentences based on an analysis of a text graph

- Sentences are vertex \((s_1, s_2, ..., s_n)\)
- There are edges \(E(s_k, s_l)\) connecting \(s_k\) with \(s_l\)
- \(\text{In}(s_i)\) is the set of sentences \(s_i\) such that there is an edge \(E(s_i, s_j)\)
- \(\text{Out}(s_i)\) is the set of sentences \(s_i\) such that there is an edge \(E(s_j, s_i)\)
- Graph generally undirected but could be directed if text order is taken into account (\(s_i\) connects with \(s_j\) only if \(i < j\))
Scientific document summarization

Document summarization overview

Summarization by sentence extraction:
graph-based techniques at document level

LexRank and TextRank score sentences based on
an iterative procedure and weighting mechanisms similar to PageRank

Sentence similarity:

- **LexRank** uses cosine similarity to compare sentences
- **TextRank** uses a kind of jaccard coefficient

✓ Parameters (d, w(s), etc.) need to be estimated
✓ Scores computed iteratively until convergence

**LexRank**

\[
  w(si) = \frac{d}{N} + (1 - d) \times \sum_{sj \in \text{In}(si)} \frac{\text{sim}(si,sj)}{\sum_{sk \in \text{Out}(sj)} \text{sim}(sk,sj)} \times w(sj)
\]

**TextRank**

\[
  w(si) = (1 - d) + d \times \sum_{sj \in \text{In}(si)} \frac{w_{j,i}}{\sum_{sk \in \text{Out}(sj)} w_{j,k}} \times w(sj)
\]

\[
  w_{j,k} = \frac{|sj \cap sk|}{\log(|sj|) + \log(|sk|)}
\]
Scientific document summarization

Document summarization overview

Summarization by sentence extraction:
other Information Retrieval Techniques

The vector space model: paragraph is represented as a vector of terms and weights
and similitude between paragraphs is computed using inner product

\[ \sim(D_i, D_j) = \sum d_{ik} \cdot d_{jk} \]

\[ D_i = (d_{i1}, \ldots, d_{in}) \]

Paragraph selection based on graph search techniques: best first, etc.

(Salton et al. 1997)
Humans identify units in ideal summaries and units in automatic summaries

Units are matched and their overlap assessed

Text quality assessed by means of questionnaires

Human evaluation is very expensive
Evaluation of automated summaries

**ROUGE: Recall-Oriented Understudy for Gisting Evaluation**

Measures content quality of a summary by comparison with ideal(s) summaries based on n-gram counting

\[
\text{ROUGE-}n = \frac{\sum_{S \in \{\text{Refs}\}} \sum_{n\text{-gram} \in S} \text{countmatch}(n\text{-gram})}{\sum_{S \in \{\text{Refs}\}} \sum_{n\text{-gram} \in S} \text{count}(n\text{-gram})}
\]

Other ROUGE metrics:

- **ROUGE-L**: Based on longest common subsequence
- **ROUGE-W**: Weighted longest common subsequence, favours consecutive matches
- **ROUGE-S**: Skip-bigram recall metric
- Arbitrary in-sequence bigrams are computed
- **ROUGE-SU** adds unigrams to **ROUGE-S**

(Lin, 2004)
Evaluation of automated summaries

Pyramid Score

based on the distribution of content units (Sus) in a set of ideal summaries, similar content units are grouped together

- each SCU in tier $T_i$ in the pyramid has weight $i$
- the best summary is one which contains all units of level $n$, then all units from $n-1$, ...
- if $D_i$ is the number of SCU in a summary which appear in $T_i$ for summary is:
  \[ D = \sum_{i=1}^{n} i \times D_i \]
- $X$ is the number of units in the summary

\[ \text{Score} = \frac{D}{\text{Max}} \]

\[ \text{Max} = \sum_{i=j+1}^{n} i \times |T_i | + j \times (X - \sum_{i=j+1}^{n} |T_i |) \]

\[ j = \max \left( \sum_{t=i}^{n} |T_t | \geq X \right) \]

(Nenkova, Passoneau, McKeown, 2007)
Document summarization overview

Summarization tools
Availability of summarization tools: no reinvent the wheel, allow comparison, provide baselines, etc.

- **MEAD**
  - publicly available toolkit for multi-lingual summarization and evaluation
  - implements different algorithms: position-based, centroid-based, it*idf, query-based summarization
  - implements evaluation methods: co-selection, relative-utility, content-based metrics

- **SUMMA**
  - publicly available: [http://www.taln.upf.edu/pages/summa.upf/](http://www.taln.upf.edu/pages/summa.upf/)
  - JAVA library to implement summarization systems
  - Statistical analysis of documents
  - Several relevance features and sentence scoring mechanisms available
  - Multilingual, Multi-document
  - Implements ROUGE and BLEU summary evaluation

(Radev et al. 2004; Saggion, 2008, 2014)
Scientific document summarization

Summarizing Scientific Articles

Scientific information overload is going to be more and more problematic. Summarization can help scientists and other interested partners to access text collections by means of automated summaries.

- Scientist and other interested parties nowadays face the problem of scientific information overload.
  - *PubMed* contains more than 24M papers, *Elsevier’ Scopus* over 57M, while *Thomson Reuther’s ISIWeb* of Knowledge more than 90M....
  - Current estimates indicate that a research paper is published every 13 seconds.

- Scientific text summarization was the first summarization application domain!
  - Summarization of scientific documents has been addressed using traditional relevance features, classification or generic/domain specific scientific information.
Scientific document summarization

Summarizing Scientific Articles

Summarization approaches to the Scientific Document
Summarization has to be adapted to the peculiarities of the scientific discourse: length, document structure, terminology, citations, rhetorical organization, etc.

• Rhetorical classification of sentences
  – Extracting sentences likely to contain semantic information on objectives, goal, own contributions, etc. ➔ classification

• Extracting scientific specific information
  – Concept based abstracting ➔ information extraction + template-based generation

• Extraction generic scientific information
  – Extracting sentences based on generic information types ➔ information extraction + shallow generation

• Relying on the opinion of the scientific community to summarize
  – Taking advantage of “citation sentences” to summarize a article
  – Impact-based summarization ➔ uses the source document
  – Citation-based summarization ➔ uses the “citation sentences”
Example in the area of crop husbandry

- SPECIES (what is studied); CULTIVAR (the variety that is studied); HIGH-LEVEL-PROPERTY (the property studied: growth); PEST (a pest that attacks the species); AGENT (the chemical/bio agent used to control the pest); etc.

Method

- Weighted patterns (PEST is a ? pest of SPECIES) are applied to the text to instantiate concepts
- Matched strings are analyzed and weighted to extract final values
- Summaries are generated using the strings

This paper studies the effect of [AGENT] on the [HLP] of [SPECIES] OR this paper studies the effect of [METHOD] on the [HLP] of [SPECIES] when it is infested by [PEST]...

This paper studies the effect of \textit{G. pallida} on the yield of \textit{potato}. An experiment in \textit{1985 and 1986} at \textit{York} was undertaken.

(Oakes and Paice, 2000)
Article Title: Features 3D scanning systems for rapid prototyping (97 sentences)

Indicative Summary: Describes two non-contact scanning systems, REVERSA and ModelMaker

Topics: CADAM system; ModelMaker; REVERSA; standard dual view system; system

Expanding Topics: REVERSA and ModelMaker

REVERSA is a dual viewpoint non-contact laser scanner which comes complete with scanning software and data manipulation tools.

The ModelMaker scanning system is a combination of a 3D laser stripe sensor, 6DOF position localizer and a PC...

ModelMaker can simply be retrofitted to existing arms providing the benefits of a portable CMM with dense depth data sets...

(Saggion and Lapalme, 2002)
Summarizing Scientific Articles

Generating Indicative-Informative Summaries of Technical Articles

Generate a brief indicative summary of the main topics discussed in the paper and expand the topics with useful information about the topics. Modelling general scientific information

• Based on a linguistic & conceptual model of the scientific article
  – Concepts = author, section, problem, solution, limitations, etc.
  – Relations = present topic, define, elaborate, conclude, etc.
  – Patterns for interpretation = dictionary elements + syntax + lexical elements
  – Templates for generation

• Text analysis: POS tagging + pattern-matching

• Sentence scoring: titles (main + section headings) & verb-argument (noun phrase) scoring guide sentence selection

• Text generation: order information based on dictionary categories, sentence fusion, verb transformation (personal (e.g. We describe X) \(\Rightarrow\) impersonal (e.g. Describes X) etc.

• Evaluation: text classification, comparison against author abstract, comparison against ideal summaries
  – Improves over all baselines
Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization

New forms of scientific summarization are based on citation networks: a paper is summarized taking into account the opinions or views of the scientific community has about a paper.


In 2014 National Institute for Standards & Technology (NIST) proposed the BioSumm Shared Task to promote the development of methods for summarizing scientific articles

Writing surveys / overviews of developments in biomedicine (or any other field) requires the analysis of considerable number of scientific publications

- **Author abstracts** do not provide information on the lasting influences of a work
- **Citations** do not provide enough context from the cited paper

Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization: BioSumm 2014 & SciSumm 2016

Structure of the dataset
- **Ech Collection** is made of 
  1 Reference Paper + 10 
  Citing Papers 
- For each Collection, **four** 
  ‘250-words’ human-written summaries of the 
  reference paper

BioSumm 2014: 30 Coll. for 
training and 20 Coll. for 
evaluation
SciSumm 2016: 10 Coll. for 
development and 10 Coll. For 
evaluation

For each Collection, three tasks are proposed:

- Task 1A: identify text spans being cited
- Task 1B: identify citation facet
- Task 2: create a community-based summary

Summarizing Scientific Articles

Citation-based summarization: fact-based citation summaries (C-LexRank)

DataSet

- ACL Anthology Network (ANN)
- 5 clusters of documents extracted (each on a given topic, matched with specific keyword e.g. dependency parsing)
- Each cluster 5 different documents, each with citations
- For each paper a “citation summary” was created based on the sentences “citing” the paper
- Annotators were asked to extract facts from the citation summary (keywords or phrases) that represent the content

(Qazvinian and Radev, 2008)
Some annotators agreed on some facts (Czech DP, lexical rules, etc.)

Some annotators found unique facts like: generative model

A \{0,1\} matrix can be created which indicates which facts are covered by which citation sentences

<table>
<thead>
<tr>
<th>Fact</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>f4: “Czech DP”</td>
<td>10</td>
</tr>
<tr>
<td>f1: “lexical rules”</td>
<td>6</td>
</tr>
<tr>
<td>f3: “POS/ tag classification”</td>
<td>6</td>
</tr>
<tr>
<td>f2: “constituency parsing”</td>
<td>5</td>
</tr>
<tr>
<td>f5: “Punctuation”</td>
<td>2</td>
</tr>
<tr>
<td>f6: “Reordering Technique”</td>
<td>2</td>
</tr>
<tr>
<td>f7: “Flat Rules”</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fact</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Dependency conversion”</td>
</tr>
<tr>
<td>“80% UAS”</td>
</tr>
<tr>
<td>“97.0% F-measure”</td>
</tr>
<tr>
<td>“Generative model”</td>
</tr>
<tr>
<td>“Relabel coordinated phrases”</td>
</tr>
<tr>
<td>“Projective trees”</td>
</tr>
<tr>
<td>“Markovization”</td>
</tr>
</tbody>
</table>

Facts for paper “A Statistical Parser for Czech” with 54 citations

The summary of a paper is created by: creating a Citation Summary Network and selecting citing sentences that cover a varied set of relevant facts

- Sentences well connected in the network (high similarity) should represent shared facts
- Different sentence similarity measures are compared to decide on the most appropriate (evaluated on paper “A Statistical Parser for Czech”)

(Qazvinian and Radev, 2008)
Network-based clustering is applied to **group sentences which share many common facts** by a hierarchical agglomerative clustering algorithm.

Evaluation is carried out computing purity where \( K \) are the clusters and \( C \) are the classes (the facts!)

\[
purity(K, C) = \frac{1}{N} \sum_{i} \max_{j} | k_{i} \cap c_{j} |
\]

**Selection of summary sentence from clusters:**

1. **Cluster Round-Robin (C-RR):** Sort the clusters by their size and extract one sentence from each cluster, then extract more sentences until compression is reached.

2. **Cluster LexRank (C-lexrank):** Inside each cluster LexRank is applied to score sentences. The most salient sentences from each cluster are selected.

**Baseline methods:**

- Random summary
- LexRank (without initial clustering)

The best performing system (according to **pyramid** scores) overall is C-lexrank, followed by Lexrank, and then by C-RR

(Qazvinian and Radev, 2008)
Impact-based summary: a set of sentences from the paper that can reflect the impact of the paper

Instead of using citation sentences the approach uses sentences from the paper (to avoid including content which is not directly related to the paper to summarize)
Citation-based summarization: Impact-driven summaries
Summarizing the impact of a scientific publication: “... the impact of a paper has to be judged based on the consent of the research community...”

- Citation context: widow of sentences around the citation
- Approach
  - Construct a representation of the impact \( I \) of a document \( d \) based on \( d \) and the citation context \( C \)
  - Develop a scoring function \( \text{Score}(.) \) to rank sentences of \( d \) reflecting \( I \)
- The approach can be seen as a retrieval problem: sentences of \( d \) are documents and \( I \) is a query: retrieve sentences matching \( I \)

(Mei and Zhai, 2008)
Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization: Impact-driven summaries

Summarizing the impact of a scientific publication: “... the impact of a paper has to be judged based on the consent of the research community...”

- **Impact Language Model**: an unigram model for I (the impact), based on both (i) the document $d$ to summarize and (ii) the citation context $C$
  - probabilities for *words in $d$* are estimated using *relative frequencies*
  - probabilities for *words in $C$* are estimated from *relative frequencies, citation paper impact* (based on page rank), and *position of the sentence* with respect to the citation marker

(Mei and Zhai, 2008)
Citation-based summarization: Impact-driven summaries

Summarizing the impact of a scientific publication: “... the impact of a paper has to be judged based on the consent of the research community...”

The scoring function of paper sentences \((\text{score}(s))\) is based on Kullback-Leibler (KL) divergence

- \(V\) is the set of words in the vocabulary
- \(\theta_I\) is the Impact Language Model
- \(\theta_s\) is the sentence language model

\[
\text{score}(s) = -D(\theta_I \parallel \theta_s) = \\
= \sum_{w \in V} p(w | \theta_s) \log( p(w | \theta_s)) - \sum_{w \in V} p(w | \theta_I) \log( p(w | \theta_I))
\]

If \(\theta_s\) and \(\theta_I\) are very close, the KL-divergence would be small and \(\text{Score}(s)\) would be high

\[(\text{Mei and Zhai, 2008})\]
Summarizing Scientific Articles

Citation-based summarization: Impact-driven summaries
Summarizing the impact of a scientific publication: “... the impact of a paper has to be judged based on the consent of the research community...”

• Data
  – SIGIR papers from 1978 to 2005 (1,303 papers)
  – Citation contexts extracted (5 sentences): sentence with citation marker -2,+2
  – Only papers with at least 20 citations are considered (14 papers)
  – Experts assessed each sentence in the paper and decided if it covers “influential” content as indicated in the citation contexts
  – The influential sentences are considered as the gold standard summaries for evaluation

(Mei and Zhai, 2008)
Summarizing Scientific Articles

Citation-based summarization: Impact-driven summaries
Summarizing the impact of a scientific publication: “... the impact of a paper has to be judged based on the consent of the research community...”

• Evaluation
  – ROUGE-1 and ROUGE-L are used to compare automatic summaries with gold summaries
  – Baselines: LEAD, MEAD, MEAD + Citation Context
  – KL-divergence summarizer outperforms all baselines
  – Parameters such as authority and proximity of sentence to citation have an impact on the results

(Mei and Zhai, 2008)
A limitation of citation-based approaches to scientific summarizations is the use of explicit citation information. Explicit citation: This approach is one of those described in Eisner (1996). Offers very little information about Eisner’s paper.

Implicit or non-explicit citation sentences may contain useful information on the cited paper.

The parser searches for the best parse for the sentence. This approach is one of those described in Eisner (1996). Non-explicit citation (Qazvinian and Radev, 2010)
Method:
Construction of a graphical model based on Markov Random Fields (MRF) from the sentences in the document

Evaluation with respect to gold-standard ($F$-measure)
Evaluation with respect to extrinsic citation-based summarization (using pyramid method)

(Qazvinian and Radev, 2010)
Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations

• For each sentence $S_i$, $C_i$ represents an event of being a non-explicit citation

• Observed nodes represent measurable information about sentences (sentence content)

• Hidden nodes represent the state of the sentence (non-explicit citation state) – modelled with a potential function $\phi(C_i)$ or probability of being at state $C_i$

• Relation between neighbouring sentences represented with a weighted edge: compatibility function $\psi_{ij}(C_j | C_i)$ represents $i$ believes about $j$

(Qazvinian and Radev, 2010)
HIDDEN NODES

OBSERVED NODES

• Assumptions about compatibility
  – if sentence is **not** “non-explicit citation” can **not** “say” much about other sentences
  – If sentence is a “non-explicit citation”, it can say something about neighbouring sentences

\[
\psi_{ij}(c_j \mid \neg c_i) = 0.5
\]

\[
\psi_{ij}(c_j \mid c_i) = S_{ij} = \frac{1}{1 + e^{-\cosine(i, j)}}
\]

(Qazvinian and Radev, 2010)
Computation of values of hidden variables (probabilities of being $c_i$) is carried out with Belief Propagation (messages are sent from one sentence to the others).

$$
\begin{align*}
    m_{ij}(c_j) &\leftarrow P(c_i)\psi_{ij}(c_j \mid c_i) \prod_{k \in ne(i) \setminus j} m_{ki}(c_i) + P(\neg c_i)\psi_{ij}(c_j \mid \neg c_i) \prod_{k \in ne(i) \setminus j} m_{ki}(\neg c_i) \\
    m_{ij}(\neg c_j) &\leftarrow P(c_i)\psi_{ij}(\neg c_j \mid c_i) \prod_{k \in ne(i) \setminus j} m_{ki}(c_i) + P(\neg c_i)\psi_{ij}(\neg c_j \mid \neg c_i) \prod_{k \in ne(i) \setminus j} m_{ki}(\neg c_i)
\end{align*}
$$

- $ne(i)$ indicates the neighbours of sentence $i$ in the network.
- Messages $m_{ij}$ are initially 0.5 and are updated through iteration (they are considered probabilities).

(Qazvinian and Radev, 2010)
Final believe values (i.e. probabilities) are computed with the final values as (with $k$ a normalization factor):

$$b(c_i) = k \phi(c_i) \prod_{j \in e(i)} m_{ij}(c_i)$$

$$b(\neg c_i) = k \phi(\neg c_i) \prod_{j \in e(i)} m_{ij}(\neg c_i)$$

Choosing a threshold for deciding if the sentence is a non-explicit citation

The values of $\phi(c_i)$ are computed with a normalized linear formula that combines

– a binary value for the presence of explicit citation
– a binary value for the presence of certain patterns
– the “cosine” similarity of the sentence to the cited paper

(Qazvinian and Radev, 2010)
Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations

- **Evaluation dataset**: set of 10 documents from the ACL anthology + their implicit citation sentences (human annotated)

- Different network configurations explored (BP₁: one neighbour, BP₄: 4 neighbours, BPₙ: all neighbours)

- Baseline systems: B₁ selects previous-following sentence if similarity greater than thr. B₂ selects any neighbouring sentences (in a 4-sentence window) matching a pattern, SVM (with 3 features) a trained model using all docs minus one for training

(Qazvinian and Radev, 2010)
Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations

- Considering 4 sentences as the context of influence provides the best results
- Network-based approach better than sentence classification
- Using implicit citations for summary generation improves results (*pyramid*) that use only explicit citations

<table>
<thead>
<tr>
<th>paper</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>SVM</th>
<th>$BP_1$</th>
<th>$BP_4$</th>
<th>$BP_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P08-2026</td>
<td>0.441</td>
<td>0.237</td>
<td>0.249</td>
<td>0.470</td>
<td>0.613</td>
<td>0.285</td>
</tr>
<tr>
<td>N07-1025</td>
<td>0.388</td>
<td>0.102</td>
<td>0.124</td>
<td>0.313</td>
<td>0.466</td>
<td>0.138</td>
</tr>
<tr>
<td>N07-3002</td>
<td>0.521</td>
<td>0.339</td>
<td>0.232</td>
<td>0.742</td>
<td>0.627</td>
<td>0.315</td>
</tr>
<tr>
<td>P06-1101</td>
<td>0.125</td>
<td>0.388</td>
<td>0.127</td>
<td>0.649</td>
<td>0.880</td>
<td>0.193</td>
</tr>
<tr>
<td>P06-1116</td>
<td>0.283</td>
<td>0.104</td>
<td>0.100</td>
<td>0.307</td>
<td>0.341</td>
<td>0.130</td>
</tr>
<tr>
<td>W06-2633</td>
<td>0.313</td>
<td>0.100</td>
<td>0.176</td>
<td>0.338</td>
<td>0.413</td>
<td>0.160</td>
</tr>
<tr>
<td>P05-1044</td>
<td>0.225</td>
<td>0.100</td>
<td>0.069</td>
<td>0.172</td>
<td>0.588</td>
<td>0.004</td>
</tr>
<tr>
<td>P05-1073</td>
<td>0.144</td>
<td>0.100</td>
<td>0.144</td>
<td>0.433</td>
<td>0.518</td>
<td>0.171</td>
</tr>
<tr>
<td>N03-1003</td>
<td>0.245</td>
<td>0.249</td>
<td>0.126</td>
<td>0.523</td>
<td>0.466</td>
<td>0.125</td>
</tr>
<tr>
<td>N03-2016</td>
<td>0.100</td>
<td>0.181</td>
<td>0.224</td>
<td>0.439</td>
<td>0.482</td>
<td>0.185</td>
</tr>
</tbody>
</table>

(Qazvinian and Radev, 2010)
Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization: improving coherence
Citations may produce incoherent summaries, so further processing might be needed

• Problems
  – Citations may contain material referring to other articles
  – Including irrelevant material will waste space
  – Ordering sentences in a citation-based summary may affect coherence/cohesion (the order may not be logical)

• Approach
  – Filtering out unsuitable citation sentences and removing irrelevant parts from citation sentences
  – Selecting best citation sentences (covering relevant aspects of the cited paper)
  – Post-process the sentences to enhance the summary

(Abu-Jbara and Radev, 2011)
Finding the **scope of the reference** is achieved by parsing the sentence and extracting the smallest sub-tree rooted $S$ (sentence) which contains the reference.

Sentences are classified as **suitable or unsuitable** using supervised learning (SVM).

Sentences are:
- **classified** as Background, Problem, Method, Result and Limitation
- in each class, **clustered** by a hierarchical agglomerative community finding algorithm
- in each cluster sentences are **weighted** using the LexRank algorithm.

**Sentences are selected** based on: their category (B, P, M, R, L), size of the cluster they belong to, and LexRank values.

Finally the sentences are **post-processed**, the citation marker can be removed or transformed into a pronominal reference (he/she/they)
- A trainable system is used to decide the appropriate transformation

(Abu-Jbara and Radev, 2011)
Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization: improving coherence

• Dataset
  – 55 papers from the ANN corpus are used
  – Citation sentences are annotated with labels: Background, Problem, Method, Result, Limitation, Unsuitable
  – Citation markers are annotated with replace, remove, or keep
• Evaluation with ROUGE-L
  – (5 sentences long) were created for 30 papers out of citation sentences
  – Baselines used: MEAD with default settings, LexRank, citation-based summaries (QV08 system previous slides)
  – System outperforms all baselines in ROUGE-L (sentence filtering having a high impact in the model)
  – System has more coherent summaries than QV08

(Abu-Jbara and Radev, 2011)
Scientific document summarization

Summarizing Scientific Articles

Citation-based summarization: generating state-of-the-art reports
A state of the art report or a survey of a scientific topic can be considered an instance of multi-document summarization

• **Automatic Related Work Summarization**
  – Combines sentences from target paper and sentences from cited papers
  – Topic tree of the state of the related work section (manually constructed)
  – Sentences attached to topic based on how well it reflects topic and a mix of author and reference papers are selected for each topic

• **Using Keywords**
  – Given an initial query (“Word Sense Disambiguation”) a precise search for paper based on matching on titles and abstracts is carried out and then expanded with papers citing/cited by the initial papers
  – For each paper citing sentences are used to generate the surveys
  – Sentences are selected based on different methods: Centroid, LexRank, and C-LexRank (clustering)
  – Pyramid scores show that best system is LexRank

(Hoang and Kan, 2010) (Jha et al., 2013)
Patent summarization

Summarizing Patents

Legal documents (US Const.)

“Art. 1, Sec. 8. The Congress shall have power . . . To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries.”

Objectives: intellectual property protection, secures markets, competitor control, etc.

Once a patent is granted knowledge is disclosed and transferred to society

Characteristics:

- long documents
- long sentences (>500W sentences are common in some sections – claims)
- complicated sentence structure (many embedded clauses and coordination)
- complicated terminology (specific classification codes, technical terms, use of peculiar references, e.g. said device, references to other patents, biblio. references, figures, drawings, measurements, chemical compounds, etc.)
- peculiar document structure (title, field of invention, abstract, prior art, claims, description, drawings, etc.)

Patent Overload!

- the European Patent Office (EPO): 90M patents 750 patent applications each day
- Derwent World Patent Index: 33M patents
- Google: 87M patents

Manually creating summaries for patents is unfeasible
Text summarizers using relevance measurement technologies and latent semantic analysis techniques provide accurate and useful summarization of the contents of text documents. Generic text summaries may be produced by ranking and extracting sentences from original documents; broad coverage of document content and decreased redundancy may simultaneously be achieved by constructing summaries from sentences that are highly ranked and different from each other. In one embodiment, conventional Information Retrieval (IR) technologies may be applied in a unique way to perform the summarization; relevance measurement, sentence selection, and term elimination may be repeated in successive iterations. In another embodiment, a singular value decomposition technique may be applied to terms-by-sentences matrix such that all the sentences from the document may be projected into the singular vector space; a text summarizer may then select sentences having the largest index values with the most important singular vectors as part of the text summary.

20 Claims, 2 Drawing Sheets

The identification stage filters the input document to determine the most important central topics. The interpretation stage clusters words and abstracts them into some encompassing concepts. Finally, the generation stage generates summaries either by outputting some parts of the input or by creating new sentences based on the interpretation of the document concepts (E. Horvitz et al., Automated Text Summarization In Summarist, in Proceedings of the TIPSTER Workshop, 1990). This generation process was not realized in the work upon which this paper was based.

The Knowledge Management (KM) system from SRA International, Inc. extracts summarization features using morphological analysis, name tagging, and co-reference resolution. The KM approach uses a machine-learning technique to determine the optimal combination of features in a particular keyword or topic search. Developing a high quality generic summarization method and system has proven very challenging. A query-relevant summary, on the other hand, presents the documents that are factually related to the initial search query; in many existing systems, creating a query-relevant summary is essentially a process of retrieving sentences from the document. It will be appreciated by those of skill in the art that this process is strongly related to the text retrieval process. Accordingly.
Long and complicated sentences

- **Content peculiarities**
  - claim **vocabulary is very vague and abstract** to obfuscate the message: [device for recording a digital information signal in an information track on a magnetic record carrier] instead of **tape recorder**
  - author abstract is also written in vague terms
  - **noun phrases are extremely long:** [device for recording .... on a magnetic record carrier]
  - a description section elaborates the claims in more concrete terms

problem for sentence extraction methods since some sentences would be overweighed by traditional methods
Summarizing Patents

Trainable patent summarization
Scoring and selection of sub-sentential units and generation of the summary based on text generation techniques: use both claims and description for selection of information

- Patent processing and text analysis
  - Segmentation of patents in text segments
  - Segmentation of each sentence
  - Mention (noun-phrases) identification based on chunking
  - Coreference resolution (adaptation of Stanford Coref. Resolution)
  - Lexical chain computation (coreference, part-whole, set membership, etc.)
  - Matching/aligning claim segments with their *descriptions*

(Codina-Filbá et al, 2015)
Patent summarization

Summarizing Patents

Trainable patent summarization

Segments are scored based on a number of classical and patent specific features

- **Mention/Lex. Chains features** (aggregated and normalized in sentences)
  - mention frequency
  - coreference chain length score
  - meronym and hyperonym chain score
  - claim relevance structure
- **Segment features**
  - best and second best similarities of segment with claims
  - length
  - is segment in claim?
  - segment mentions the patent invention?
- **Classical features**
  - similarity to author summary
  - similarity to patent title
  - similarity to claims
  - tf*idf score for segment based on statistics for claims, description, abstract

(Codina-Filbá et al, 2015)
Summarizing Patents

Trainable patent summarization

Scorer is implemented with linear regression where weights are adjusted with training data

- **Data**
  - 26,498 sentences scored based on their similarity to an ideal abstract

- **WEKA linear regression (LR) used to learn optimal weights**

- **SUMMA used to implement features, compute, and select segments**

<table>
<thead>
<tr>
<th>#</th>
<th>description</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>mention frequency</td>
<td>0.1842</td>
</tr>
<tr>
<td>2.</td>
<td>coreference chain score</td>
<td>0.0665</td>
</tr>
<tr>
<td>3.</td>
<td>meronym/bolonym and hyponym/hyperonym chain score</td>
<td>0.2270</td>
</tr>
<tr>
<td>4.</td>
<td>claim structure relevance</td>
<td>0.0202</td>
</tr>
<tr>
<td>5.</td>
<td>best segment alignment similarity</td>
<td>-0.0068</td>
</tr>
<tr>
<td>6.</td>
<td>second best segment alignment similarity</td>
<td>0.0250</td>
</tr>
<tr>
<td>7.</td>
<td>segment length relevance</td>
<td>0.0143</td>
</tr>
<tr>
<td>8.</td>
<td>segment position relevance in claims</td>
<td>0.0000</td>
</tr>
<tr>
<td>9.</td>
<td>segment position relevance in background</td>
<td>0.0498</td>
</tr>
<tr>
<td>10.</td>
<td>segment position relevance in drawings</td>
<td>-0.0318</td>
</tr>
<tr>
<td>11.</td>
<td>segment position relevance in embodiment</td>
<td>-0.0214</td>
</tr>
<tr>
<td>12.</td>
<td>segment position relevance in summary</td>
<td>-0.0265</td>
</tr>
<tr>
<td>13.</td>
<td>invention segment</td>
<td>0.2830</td>
</tr>
<tr>
<td>14.</td>
<td>similarity to the summary</td>
<td>0.6025</td>
</tr>
<tr>
<td>15.</td>
<td>similarity to title</td>
<td>0.1597</td>
</tr>
<tr>
<td>16.</td>
<td>similarity to claims</td>
<td>0.0000</td>
</tr>
<tr>
<td>17.</td>
<td>mention distribution in claims</td>
<td>-0.3397</td>
</tr>
<tr>
<td>18.</td>
<td>mention distribution in abstract</td>
<td>-0.2544</td>
</tr>
<tr>
<td>19.</td>
<td>mention distribution in description</td>
<td>0.5101</td>
</tr>
</tbody>
</table>

Good predictive power of the LR model
Most features correlate well with relevance
Patent summarization

Summarizing Patents

Trainable patent summarization

Generate an abstract based on the content units selected

- Complete units and adjust grammar
- Remove parts of segments or drop segments
- Increase cohesion:

\[
\text{[a device (for) containing a signal processing unit]} \\
\text{[a device to contain a signal processing unit]} \\
\text{[a device which contains a signal processing unit]} \\
\Rightarrow \text{A device contains a signal processing unit}
\]

\[
\text{[a unit contained in a rectangular device]} \Rightarrow \text{A unit is contained in a rectangular device}
\]

\[
\text{[a sail for a sailboard]}_{\text{invention} = \text{initial}} \\
\text{[a device for coupling a sail batten to a mast in a board]}_{\text{sail}} \\
\text{[a first end for rotateably bearing against a mast]}_{\text{component} = \text{yes}} \\
\Rightarrow \text{What is claimed is a sail for a sailboard. The invention covers a device for coupling a sail batten to a mast in a board sail. The device contains a first end for rotateably bearing against a mast.}
\]

- Content evaluation: mention recall, precision, f-measure \(\Rightarrow\) system outperforms LexRank, Centroid, and LEAD
- Human content evaluation: similar results
Conclusions

• The information provided by citations is essential to support and improve the generation of summaries of scientific documents.
• Several kinds of information can be included in a summary of a scientific publication: the relevant contents of the paper, which parts of the papers had more impact on the research community, the feedback of the research community concerning a specific article, etc.
• Multi-document summarization is useful to help the creation of state-of-the-art reports.
• General purpose metrics and techniques have to be adapted in order to assess scientific content.
Cited works (1/3)

Amjad Abu-Jbara, Dragomir R. Radev. Coherent Citation-Based Summarization of Scientific Papers. ACL 2011: 500-509
Horacio Saggion. Creating Summarization Systems with SUMMA. LREC 2014: 4157-4163
Horacio Saggion. SUMMA. A Robust and Adaptable Summarization Tool. TAL 49(2): 103-125 (2008)
Scientific document summarization

Cited works (2/3)


Lin, CY. ROUGE: A Package for Automatic Evaluation of summaries. ACL Summarization Workshop 2004

Mead et al. MEAD - a platform for multidocument multilingual text summarization. In LREC 2004.


Oakes, M. and Paice, C.D. Term extraction for automatic abstracting.


CHALLENGES, DATASETS AND ARCHITECTURES
Outline

• Scientific Literature Mining Challenges
• Datasets and tools
• Structured / semantic publication formats
• Scholarly literature architecture
Several challenges have been organized to explore how we can take advantage of scientific literature to automatically carry out specific text analysis tasks.
Scientific Literature Mining Challenges

**Issue**: author-name ambiguity (authors that publish with several name variations, different authors sharing the same name)

**Dataset** (from *Microsoft Academic Graph*):
- 250k (authors + affiliation)
- 2,5M (papers + conference / journal info)
- Author/paper pairs (to evaluate if correct or not) ground truth on manual corrections of Microsoft Academic Graph

**Two Tracks**:
1. **Author-Paper identification**: for each author papers that she has written
2. **Author disambiguation challenge**: group duplicated author names referring to the same author

Track 1: extensive feature engineering on the MAG and binary classifier of paper-author pairs
Track 2: multi step approach for string name processing and matching

None of these approaches directly scales sufficiently well for use on the entire Microsoft Academic Search author and publication data

Scientific Literature Mining Challenges

**KDD Cup 2016**

**Issue**: given a research field, rank the relevance of institutions

**Dataset**: any dataset publicly available online together with the *Microsoft Academic Graph* can be used

**Track**: Rank a set of institutions with respect to the number of full research papers they get accepted in 2016 conferences: SIGIR, SIGMOD, SIGCOMM, KDD, ICML, FSE, MobiCom, MM

Great predictive power of the participation of the institution in the past editions of the conference

Challenges, datasets and architectures

Scientific Literature Mining Challenges

**Issue:** assess the query-independent importance of scholarly articles

**Dataset:** any dataset publicly available online together with the *Microsoft Academic Graph* can be used

**Track:**
Generate static ranking of papers with respect to their relevance

Iterative solution that refine citation-graph paper ranking measures by means of the information concerning paper authors and venue of publication

Challenges, datasets and architectures

Scientific Literature Mining Challenges

Semantic Publishing Challenge

2014 / 2015 / 2016 – in conjunction with Extended Semantic Web Conference

Issue: automatically generate semantic publishing RDF datasets from both conference proceedings and papers

Dataset:
CEUR-WS Web proceedings (task 1), CEUR-WS papers in PDF format (task 2), RDF semantic publishing datasets (task 3)

Tasks (2016):
1. **Extract information from CEUR-WS online proceeding (HTML)** (what workshop series a workshop is part of, affiliations of editors, exact date of workshop and of proceedings publication, distinction between invited and contributed papers)
2. **Extract information from PDF files of papers from CEUR-WS** (author, affiliations and countries, captions of tables and figures, funding agencies, EU projects, sections)
3. **Interlink semantic publishing RDF datasets**

Challenges, datasets and architectures

Scientific Literature Mining Challenges

Semantic Publishing Challenge

2014 / 2015 / 2016 – in conjunction with Extended Semantic Web Conference

Task 1

1. Linguistic & structural analyzer
2. Semantic annotator
3. Annotation sanitizer
4. External resources linker
5. RDF generator

Task 2

1. PDF-to-Text converter
2. Linguistic & structural analyzer
3. Biblio parser
4. Spot ontology / founding body mentions
5. RDF generator

RDF data model

SPARQL evaluation query

WHERE {
  ?proceedings a swrc:Proceedings ;
  swrc:homepage ?procURL ;
  swrc:isAbout ?workshop .
  ?workshop a swrc:Workshop ;
  swrc:atEvent ?paper .
  ?paper a swrc:InProceedings ;
  swrc:pages ?numPages
} GROUP BY ?procURL

Number and avg. page number of papers in each proceeding

Endogenous small RNAs (miRNAs) regulate gene expression by mechanisms conserved across metazoans. While the number of verified human miRNAs is still expanding, only few have been functionally annotated.

**References**

Challenges, datasets and architectures

Scientific Literature Mining Challenges

**SemEval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles**

- 100 articles for training and 144 for testing (from ACM Digital Library)
- converted by pdftotext
- keyphrases present in the text of the papers identified by authors and students

**SemEval-2017 Task 10: ScientceIE - Extracting Keyphrases and Relations from Scientific Publications**

- Corpus: Science Direct, 500 journal articles evenly distributed among the domains Computer Science, Material Sciences and Physics
- training: 350 documents, development: 50 documents, test: 100 documents
  - **task 1**: Identification of keyphrases
  - **task 2**: Classification of identified keyphrases (PROCESS, TASK and MATERIAL)
  - **task 3**: identification of relations among keyphrases: HYPONYM-OF, SYNONYM-OF, NONE

Information extraction is the process of extracting structured data from unstructured text, which is relevant for several end-to-end tasks, including question answering. This paper addresses the tasks of named entity recognition (NER), a subtask of information extraction, using conditional random fields (CRF). Our method is evaluated on the ConLL-2003 NER corpus.
Datasets and tools

The ACL anthology network corpus

- last release: December 2013
- PDFbox to convert PDF papers
- semi-automated manual editing

<table>
<thead>
<tr>
<th>Number of papers</th>
<th>21,212</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of authors</td>
<td>17,792</td>
</tr>
<tr>
<td>Number of venues</td>
<td>342</td>
</tr>
<tr>
<td>Number of paper citations</td>
<td>110,976</td>
</tr>
<tr>
<td>Number of author collaborations</td>
<td>142,450</td>
</tr>
<tr>
<td>Citation network diameter</td>
<td>22</td>
</tr>
<tr>
<td>Collaboration network diameter</td>
<td>15</td>
</tr>
</tbody>
</table>

ACL Anthology SearchBench

- last update: November 2013, 28,000 papers
- commercial OCR to parse PDF
- integrates CiBRO to visualize citation network

<table>
<thead>
<tr>
<th>Statements</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain Text</td>
<td>+</td>
</tr>
<tr>
<td>Extracted Topics</td>
<td>+</td>
</tr>
<tr>
<td>Publication</td>
<td>+</td>
</tr>
<tr>
<td>Authors</td>
<td>+</td>
</tr>
<tr>
<td>Year</td>
<td>+</td>
</tr>
<tr>
<td>Title</td>
<td>+</td>
</tr>
<tr>
<td>Affiliations</td>
<td>+</td>
</tr>
<tr>
<td>Affiliation Sites</td>
<td>+</td>
</tr>
</tbody>
</table>
Datasets and tools

- open access content aggregator
- 37,634,579 papers with bibliographic record + PDF
- 6000 journals, collected from over 2300 Open Access repositories around the world (OAI-PMH)
- Web API
- metadata: authors, abstract, topics, year, provided by OAI

- more than 130,000,000 researcher profiles and 100,000,000 papers from multiple publication databases
- Services: Researcher profile extraction (connection with social networks like LinkedIn and VideoLectures), expert finding, social network search, topic browser, conference analysis
- Web API
Datasets and tools

• open access digital library search engine (all docs with full text)
• extract and index both metadata and full text
• provides access to metadata by means of OAI
• index also tables and figures
• 20,000 to 40,000 new crawled PDF per day – 10 PDF downloaded per second

Computer science and Neuroscience papers from: ArXiv, DBLP, CiteSeer, OdySci Academic, Aminer
• cits. count estimated (statistical model)
• keyphrases
• citation velocity and acceleration
• influential authors

CiteSeer\(^x\) 10M
http://citeseerx.ist.psu.edu/

Semantic Scholar
https://www.semanticscholar.org/
Even if 80% of scientific literature is accessed as PDF documents, structured textual formats to model the contents of scientific publications are increasingly spreading.

- **JATS XML**: an de facto standard for archiving and interchange of scientific open-access journals and its contents with XML
- **Major publishers have their own XML schemas**: Elsevier, Springer

**Semantic Web and Scholarly data**

Set of ontologies that support the creation of comprehensive machine-readable RDF metadata for every aspect of semantic publishing and referencing

- FRBR-aligned Bibliographic Ontology (FaBiO)
- Citation Typing Ontology (CiTO)
- Bibliographic Reference Ontology (BiRO)
- Citation Counting and Context Characterisation Ontology (C4O)
- **Document Components Ontology (DoCO)**
- **Publishing Status Ontology (PSO)**
- **Publishing Roles Ontology (PRO)**
- **Publishing Workflow Ontology (PWO)**
- **Discourse Elements Ontology (DEO)**
Scholarly literature architectures

Mendeley Suggest

Provide users with articles that help them to keep up-to-date with research in their field and explore relevant research that is, as of yet, unknown to them.

Multi-model recommendation: collaborative filtering (Mahout), popularity-based model, trending-based model, content-based model.

Filter out data not needed by the recommender model.

Online modules to adapt dynamically recommendations to user interactions.

Consolidate the consistency of the list of recommendation.

Online models to adapt dynamically recommendations to user interactions.

Event service: Track user interactions

Article service: crowdsourced collection of Mendeley articles

Profile service: user profile information (academic discipline, etc.)
Challenges, datasets and architectures

Scholarly literature architectures

**CiteSeer**

1. **Academic and non-academic classification**: SVM – features: document length, inclusion of bibliography, etc.
2. **Paper de-duplication**:
   - Exact PDF match: SHA1 digest
   - Near-duplicate match: based on document signature strings
3. **Metadata extraction**:
   - Header: cascade of SVM classifiers
   - Body
   - Citations: ParsCit (CRF-based)
4. **Author name disambiguation**: author names grouped into blocks of similar names. Names are matched by comparing features like titles of edited papers, co-authors, etc.

40,000 lines of codes
10 person-years for the development
[https://github.com/SeerLabs/CiteSeerX](https://github.com/SeerLabs/CiteSeerX)

**Crawler**:  "python" "django"

**Name disambiguation**:  "rails"
DR. INVENTOR SCIENTIFIC TEXT MINING FRAMEWORK

http://drinventor.eu/
FP7 ICT 2013.8.1, Grant no.: 611383
Outline

• Dealing with scientific articles in Dr. Inventor
• Dr. Inventor Text Mining Framework
  – Architectural overview
  – Hands-on Dr. Inventor Framework
Dealing with scientific articles in Dr. Inventor

The (bootstrap of) textual analyses of scientific publications often still constitutes a **time-consuming activity** due to:

- heterogeneous **input formats** (PDF, XML schemas, etc.)
- lack of **explicit structural and semantic information**
- need to **enrich contents by leveraging on external data sources**
- lack of **convenient facilities to easily access and process contents**
Dealing with scientific articles in Dr. Inventor

A scientific information mining infrastructure useful to:

- analyze publications and track research topics
- assess the novelty of ideas
- stimulate researchers' creativity by suggesting analogies between scientific outcomes

http://drinventor.eu/

FP7 ICT 2013.8.1, Grant no.: 611383
Dr. Inventor Text Mining Framework

- Integrate and customize **text mining tools** and **on-line services** to enable and ease a wide range of scientific publication analyses.
- Papers are enriched with **structural, linguistic and semantic information**.

**http://driframework.readthedocs.io/**

- Self-contained **Java 8** library managed by **Maven™**.
- Focused on **textual content**.
- Relying on a **shared data model** (java classes) to represent a paper.
- Exposing a **convenient API** to access the mined information.
- Based on **GATE** to manage **textual annotations**.
Architectural overview

Dr. Inventor Text Mining Framework

- PDF to text converter
- Inline citation spotter
- Sentence splitter
- Web based reference parser
- Citation-aware dep. parser
- Rhetorical annotator
- Babelfy WSD and Entity Linker
- Coreference res., causality spotter & graph builder
- Extractive summarizer


1. identification of inline citation markers and spans → JAPE rules
2. linking of inline citation markers to bibliographic entries
3. identification of syntactic / non-syntactic role of inline citation spans

Bibliography

Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. Both (Marcu, 2002) and (Zhang, 2003) consider a sentence pair as different realizations of a sequence of concepts. These alignment approaches segment the sentences into a sequence of phrases.

Customization of ANNIIE sentence splitter
Rule set adapted to peculiarities of scientific papers by analyzing the **most frequent sentence split patterns / errors** in a set of 40 Computer Graphics papers
Based on **MATE dependency parser**

Inline citation spans should be considered as a word-token by the parser if they have a **syntactic role**: 

```
demonstrated that this method can be improved .
```
Architectural overview

Dr. Inventor Text Mining Framework

- PDF to text converter
- Inline citation spotter
- Sentence splitter
- Web based reference parser
- Citation-aware dep. parser
- Rhetorical annotator
- Babelfy WSD and Entity Linker
- Coreference res., causality spotter & graph builder
- Extractive summarizer

Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. Both (Marcu, 2002) and (Zhang, 2003) consider a sentence pair as different realizations of a sequence of concepts. These alignment approaches segment the sentences into a sequence of phrases.

- **Rhetorical categories**: Background, Challenge, Approach, Outcome, Future Work
- **Linguistic and syntactic sentence features** exploited to train Logistic Regression classifier on Dr. Inventor Corpus (40 Computer Graphics papers including 8,777 sentences)
Kan et al. (2002) use annotated bibliographies to cover certain aspects of summarization and suggest using metadata and critical document features as well as the prominent content-based features to summarize documents. Kupiec et al. (1995) use a statistical method and show how extracts can be used to create summaries but use no annotated metadata in summarization.
These modules undergo deformation caused by force fields. They also implement attractive and repulsive force fields.
Architectural overview

Dr. Inventor Text Mining Framework

- PDF to text converter
- Inline citation spotter
- Sentence splitter
- Web-based reference parser
- Citation-aware dep. parser
- Rhetorical annotator
- Babelfy WSD and Entity Linker
- Coreference res., causality spotter & graph builder
- Extractive summarizer

Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. Both [Marcu, 2002] and [Zhang, 2003] consider a sentence pair as different realizations of a sequence of concepts. These alignment approaches segment the sentences into a sequence of phrases.

Summary:
Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. These alignment approaches segment the sentences into a sequence of phrases.

- Based on SUMMA text summarization toolkit
- Sentence relevance ranking approaches:
  - TF-IDF centroid of each section
  - TF-IDF similarity with title
  - LexRank / TextRank (soon)
  - LR of sentence features (soon)
Hands-on Dr. Inventor Framework

Full documentation and examples at: http://driframework.readthedocs.io/

- Lazy loading
- Object caching
- Factory Design Pattern to manage resource allocations

**Importing the library**

**MAVEN:**

```xml
<repositories>
  <repository>
    <id>backingdata-repo</id>
    <name>Backingdata repository</name>
    <url>http://backingdata.org/dri/library/mavenRepo/</url>
  </repository>
</repositories>

<dependency>
  <groupId>edu.upf.taln.dri</groupId>
  <artifactId>lib</artifactId>
  <version>1.0</version>
</dependency>
```

**JAVA:**

Download ZIP file with JAR and dependencies
Configure programmatically the library

Which PDF-to-text converter?
// To use PDFX:
Factory.setPDFtoTextConverter(PDFtoTextConvMethod.PDFX);

// To use GROBID:
Factory.setPDFtoTextConverter(PDFtoTextConvMethod.GROBID);

Which modules are enabled?
// Instantiate the ModuleConfig class - the constructor sets all modules 
// enabled by default
ModuleConfig modConfigurationObj = new ModuleConfig();

// Disable the parsing of bibliographic entries by means of online 
// services (Bibsonomy, CrossRef, FreeCite)
modConfigurationObj.setEnableBibEntryParsing(false);

// Disable the association of a rhetorical category to the sentences of 
// the paper
modConfigurationObj.setEnableRhetoricalClassification(false);

// Import the configuration parameters set in the ModuleConfig instance
Factory.setModuleConfig(modConfigurationObj);
Hands-on Dr. Inventor Framework

Full documentation and examples at: [http://driframework.readthedocs.io/](http://driframework.readthedocs.io/)

### Import PDF / JATS XML from file / URL

```java
// From File (substitute parsePDF for parseJATS to import JATS file):
Document doc_PDFfile = Factory.getPDFloader().parsePDF("/my/file/path/PDF_file_name.pdf");

// From URL (substitute parsePDF for parseJATS to import JATS file):
```

### Get ordered lists of sentences

```java
// Only abstract sentences
List<Sentence> abstract_SentList = doc_PDFfile.extractSentences(SentenceSelectorENUM.ONLY_ABSTRACT);

// Only body sentences
List<Sentence> abstract_SentList = doc_PDFfile.extractSentences(SentenceSelectorENUM.ALL_EXCEPT_ABSTRACT);

// Only abstract sentences
List<Sentence> abstract_SentList = doc_PDFfile.extractSentences(SentenceSelectorENUM.ALL);
```
Print the content of the first sentence of the abstract (by the asString method of the Sentence object instance)

```java
// Get ordered list of abstract sentences
List<Sentence> abstract_SentList =
    doc_PDFfile.extractSentences(SentenceSelectorENUM.ONLY_ABSTRACT);

// Get the first sentence of the abstract
Sentence firstAbstractSentence = abstract_SentList.get(0);

// Print all the data associated to the first sentence of the abstract
System.out.println(firstAbstractSentence.asString(true));
```

[SENTENCE] ID: '22047',
Text: 'Puppetry has been a popular art form for many centuries in different cultures, which becomes a valuable and fascinating heritage assert.',
Rhetorical class: 'DRI_Background'
23 TOKENS ASSOCIATED


Hands-on Dr. Inventor Framework

Full documentation and examples at: http://driframework.readthedocs.io/

Print the content of the paper sections in document order

```java
// Get ordered list of document sections
List<Section> sectionList = doc_PDFfile.extractSections(false);
for (Section sec : rootSectionList) {
    // Print all the data associated to the section
    System.out.println(sec.asString(true));
    // Get the list of sub-sections
    List<Section> subSection = sec.getSubsections();
    // Get the list of sentences inside the section
    List<Sentence> sentencesOfSection = sec.getSentences();
}
```

[SECTION] ID: '21452', Name: '1. INTRODUCTION', Level: '1', Children sections IDs: '[]', Sentences IDs: '[22053, 22054, 22055, 22056, 22057, 22058, 22059, 22060, 22061, 22062, 22063]

[SECTION] ID: '21453', Name: '2. RELATED WORK', Level: '1', Children sections IDs: '[21454, 21455]', Sentences IDs: '[]'


[SECTION] ID: '21455', Name: '2.2 Swept Surface Modelling', Level: '2', Children sections IDs: '[]', Sentences IDs: '[22082, 22083, 22084, 22085, 22086, 22076, 22077, 22078, 22079, 22080, 22081]' ...
Hands-on Dr. Inventor Framework

Full documentation and examples at: http://driframework.readthedocs.io/

Print the content of the bibliographic entries / citations

```java
// Get ordered list of bibliographic entries
List<Citation> citations = doc_PDFfile.extractCitations();

for(Citation citation : citations) {
    // Print all the data associated to the citation
    System.out.println(citation.asString(true));
}
```

Text: 'Realtime performance-based facial animation T Weise S Bouaziz H Li Pauly M ACM Transactions on Graphics (TOG) 30 4 77'

[AUTHOR] Full name: ', Thibaut Weise', First name: ', Thibaut', Surname: 'Weise'
[AUTHOR] Full name: ', Hao Li', First name: ', Hao', Surname: 'Li'
[AUTHOR] Full name: ', Mark Pauly', First name: ', Mark', Surname: 'Pauly'
PUB ID TYPE: DOI - VALUE: http://dx.doi.org/10.1145/2010324.1964972

[CIT MARKER] ID: '40228', Citation ID: '21440', Sentence ID: '22067', Reference text: '10'
[CIT MARKER] ID: '40241', Citation ID: '21440', Sentence ID: '22071', Reference text: '10'
Several high level tasks look for either one-way rewriting between single sentences, like recognizing textual entailment (RTE) (Dagan et al., 2006), or two-way rewritings like paraphrase identification (Dolan et al., 2004) and semantic textual similarity (Agirre et al., 2012).

Our system based on type-enriched string rewriting kernels obtains state-of-the-art results on paraphrase identification and answer sentence selection and outperforms comparable methods on RTE.

String rewriting kernels (Bu et al., 2012) count the number of common rewritings between two pairs of sentences seen as sequences of words.

Following the terminology of string kernels, we use the term string and character instead of sentence and word.

A type-enriched string rewriting kernel (TESRK) is simply a string rewriting kernel as defined in Equation 1 but with R a set of typed rewriting rules.

However, it cannot match the pair of sentences (C) in the original kb-SRK.

We experimented on three tasks: paraphrase identification, recognizing textual entailment and answer sentence selection.

Recognizing Textual Entailment asks whether the meaning of a sentence hypothesis can be inferred by reading a sentence text.

A SVM classifier with this kernel yields state-of-the-art results in paraphrase identification and answer sentence selection and outperforms comparable systems in recognizing textual entailment.
Save and reload a processed paper - serialized as XML

// Get the raw text contents of the paper
String rawText = doc_PDFfile.getRawText();

// Get the XML serialization of the contents of the paper, including
// all the metadata already extractor
String XMLText = doc_PDFfile.getXMLString();

// Save the XML serialization of the contents of the paper to
// the file: /my_path/stored_paper.xml
...

// Reload the contents of the paper from
// the file: /my_path/stored_paper.xml
Document doc_PDFfile_Loaded = Factory.createNewDocument("/my_path/stored_paper.xml");
GLOBAL CONCLUSIONS
AND DISCUSSION

• There is considerable room for improvement in the next future with respect to the automation of the analysis, aggregation and summarization of scientific literature

• The Natural Language Processing community plays a key role in providing better automated techniques to mine scientific literature

• The investigation of scientific text mining approaches should take into account both their effectiveness and the possibility to scale over large, heterogeneous and dynamic collections of papers
Natural Language Processing
for Intelligent Access
to Scientific Information

Many thanks for your attention!

Francesco Ronzano and Horacio Saggion

Natural Language Processing Group (TALN)
Universitat Pompeu Fabra, Barcelona, Spain

Tutorial @ COLING 2016
11th December 2016