Scholarly Data Mining: Making Sense of Scientific Literature

Horacio Saggion & Francesco Ronzano

Natural Language Processing Group (TALN)

Universitat Pompeu Fabra, Barcelona, Spain

Tutorial @ JCDL 2017

19th June 2017

TALN @ UPF

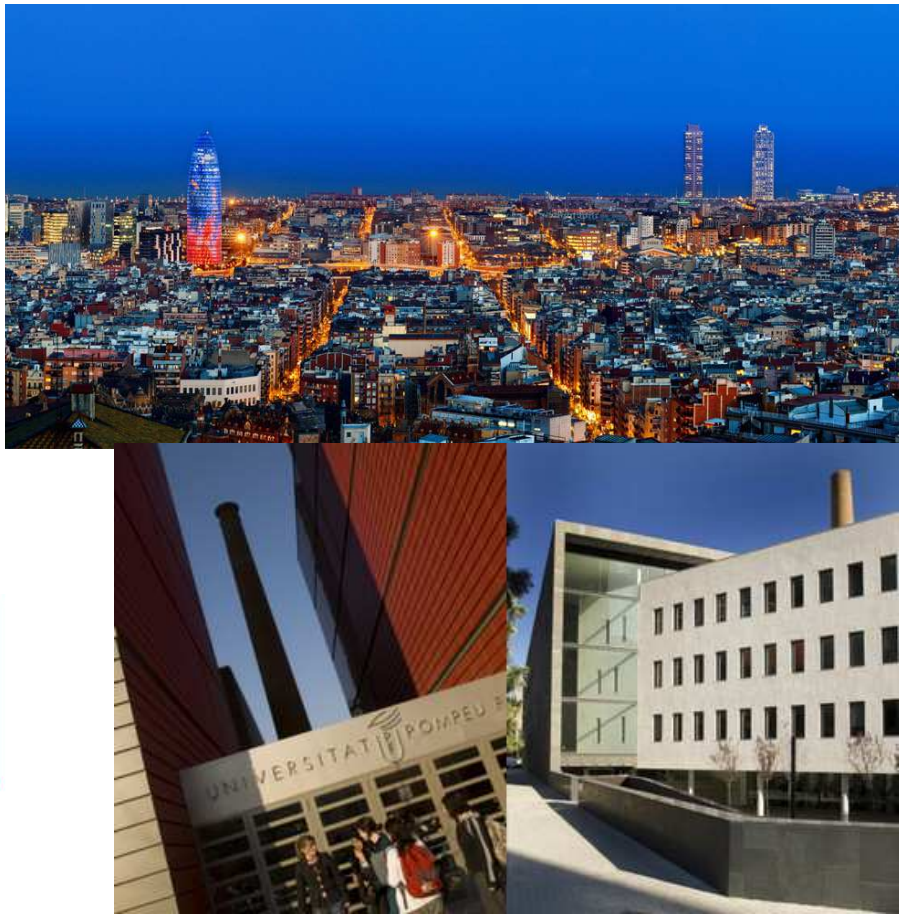
upf. Universitat Pompeu Fabra Barcelona

Department of Information and Communication Technologies

EXCELENCIA MARÍA DE MAEZTU

upf. Universitat Pompeu Fabra Barcelona

taln



- ***Horacio Saggion***, Associate Professor @ DTIC
- Francesco Ronzano, Senior Post-doc Researcher @ GRIB

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

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 AND DATASETS**
  - Which datasets are available for scientific text mining? Which tasks have been proposed?
- **DR. INVENTOR TEXT MINING FRAMEWORK**
  - Whic scientific data analyses are supported? How can the framework be used in practice?
- **GLOBAL CONCLUSIONS AND DISCUSSION**

upf. Universitat Pompeu Fabra Barcelona

taln

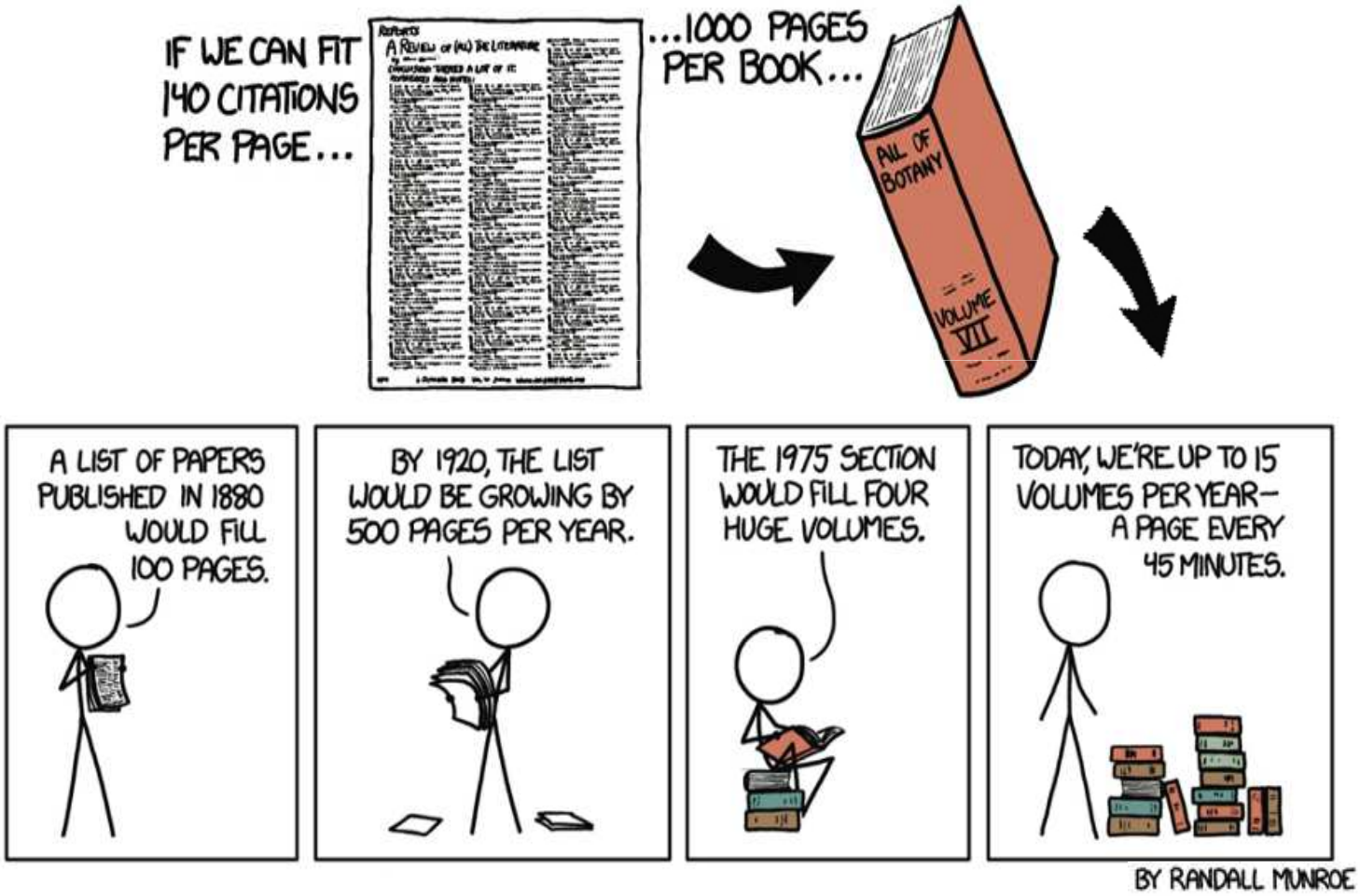
EXCELENCIA MARÍA DE MAEZTU

SCIENTIFIC INFORMATION OVERLOAD

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



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,ooo journals

Google scholar

between 100 and 160 million docs

(journal articles, books and grey literature, etc.)

PubMed

26 million publications

Journal Citation Reports
THOMSON REUTERS

11,365 journals

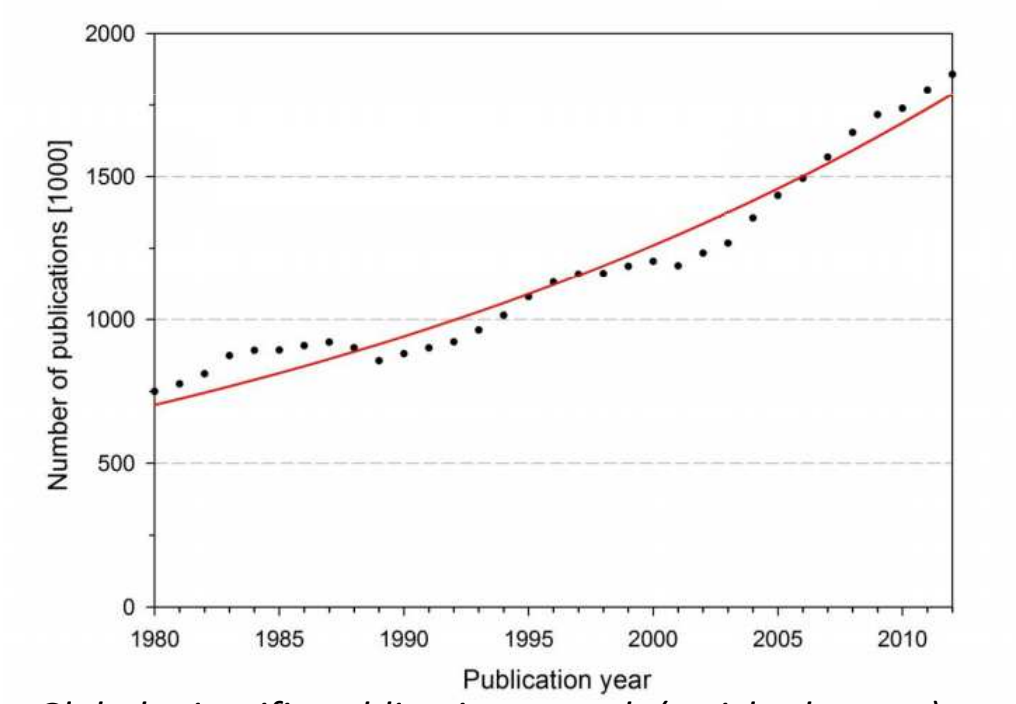
from more than 2,600 publishers

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)



Global scientific publication growth (articles by year)

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)

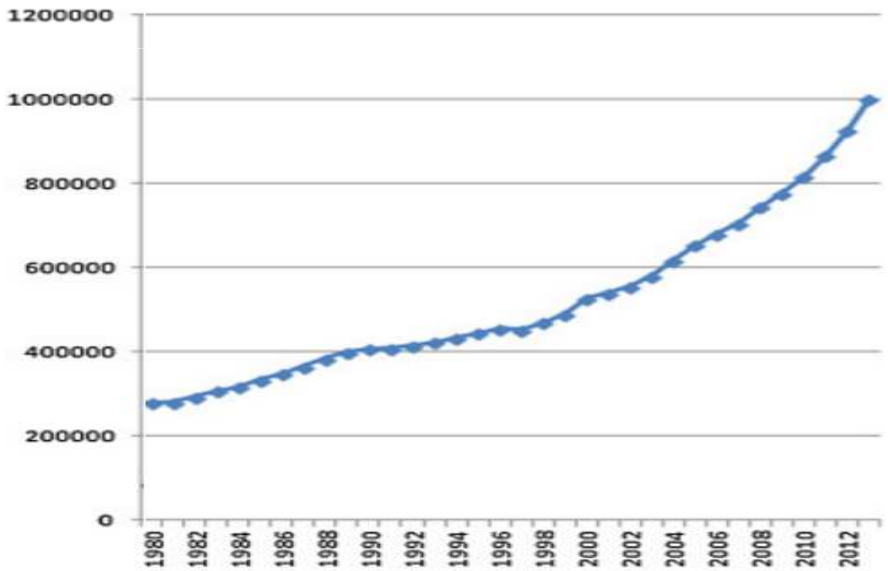
STM Report 2015

Bornmann & Mutz (2015). Growth rates of modern science.

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)



PubMed growth (articles by year)

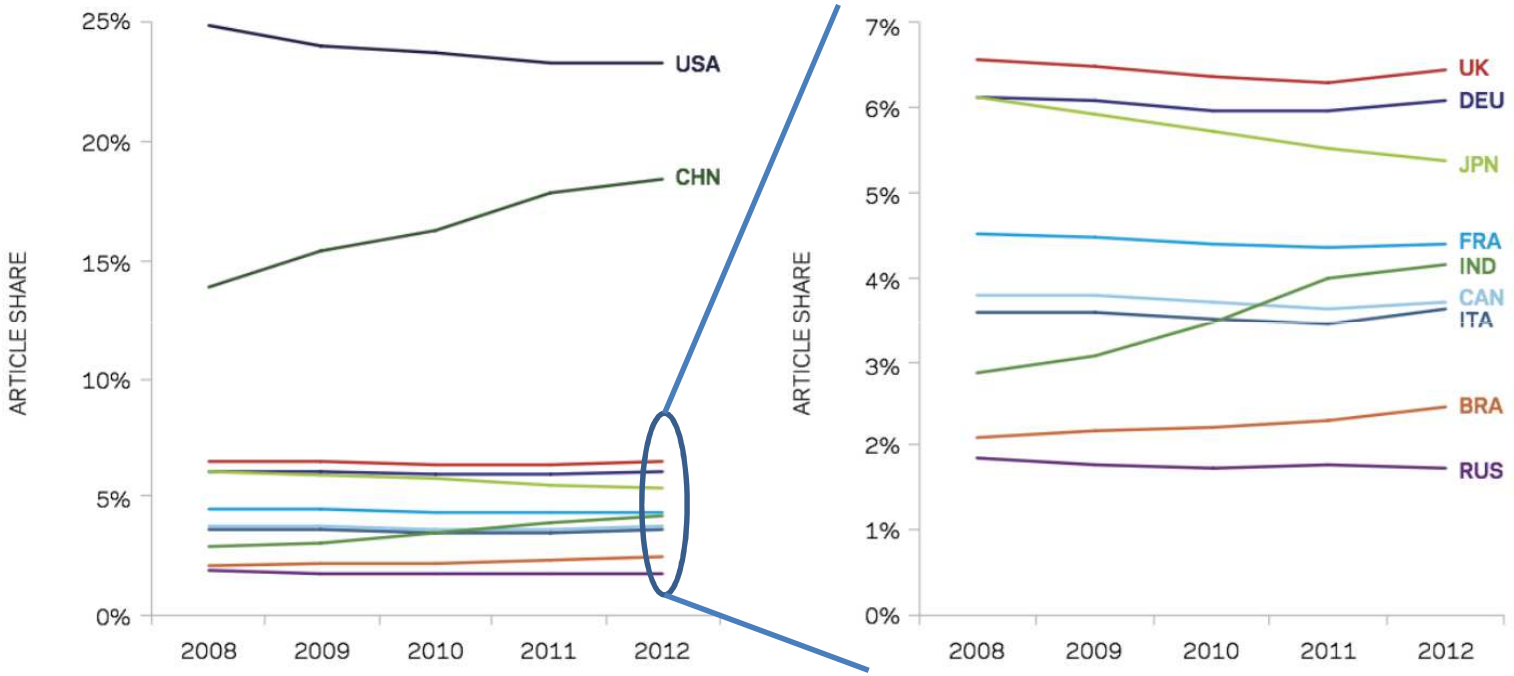
PubMed: from 1980 to 2003 the average growth is 2,9% per year, while from 2003 to 2013 it raised up to 6,7% per year

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%.

STM Report 2015 / PubMed

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

Elsevier, 2013

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

STM Report 2015

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

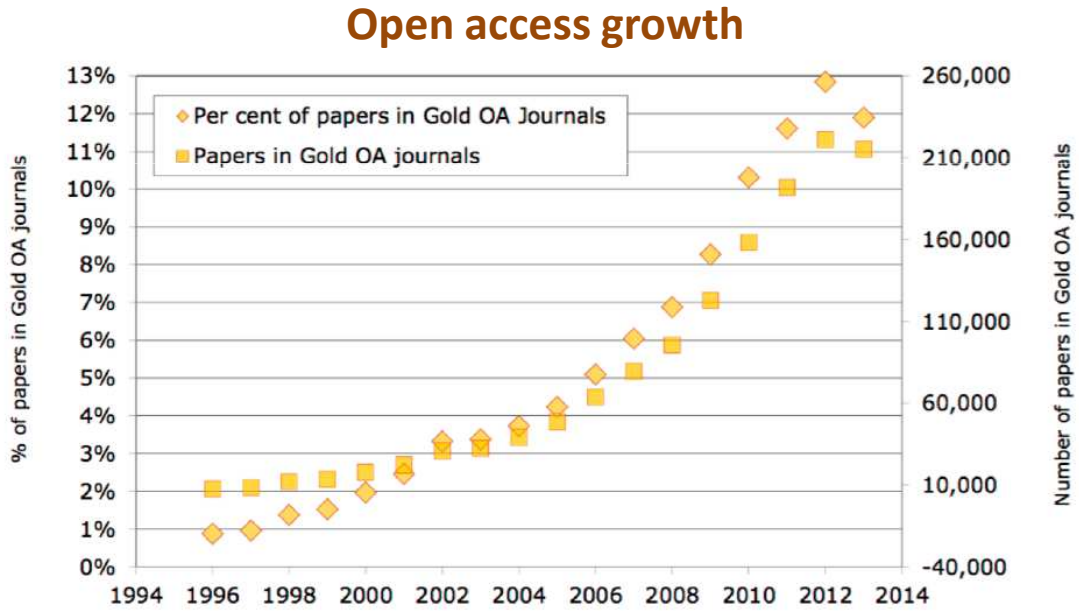
DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

9,237 journals, 2,330,000 articles

Archambault et al. (2014). Proportion of open access papers published in peer-reviewed journals at the European and world levels—1996–2013.
Lewis, D. W. (2012). The inevitability of open access.

How ‘open’ is scientific literature?

The Open Access publishing model is consistently growing



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

Archambault et al. (2014). Proportion of open access papers published in peer-reviewed journals at the European and world levels—1996–2013.

Lewis, D. W. (2012). The inevitability of open access.

Scientific Literature Overload

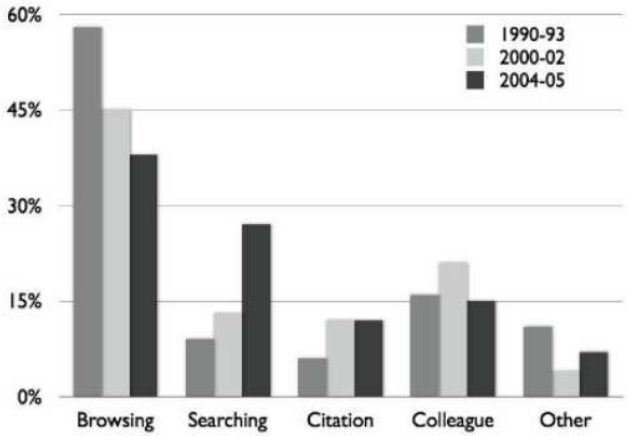
How researchers search and read publications?

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 form one search engine, Google Scholar
- Publisher sites are accessed at article level (reduced importance of publisher site browsing)

STM Report 2015
Tenopir (2007) What does usage data tell us about our users?

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:

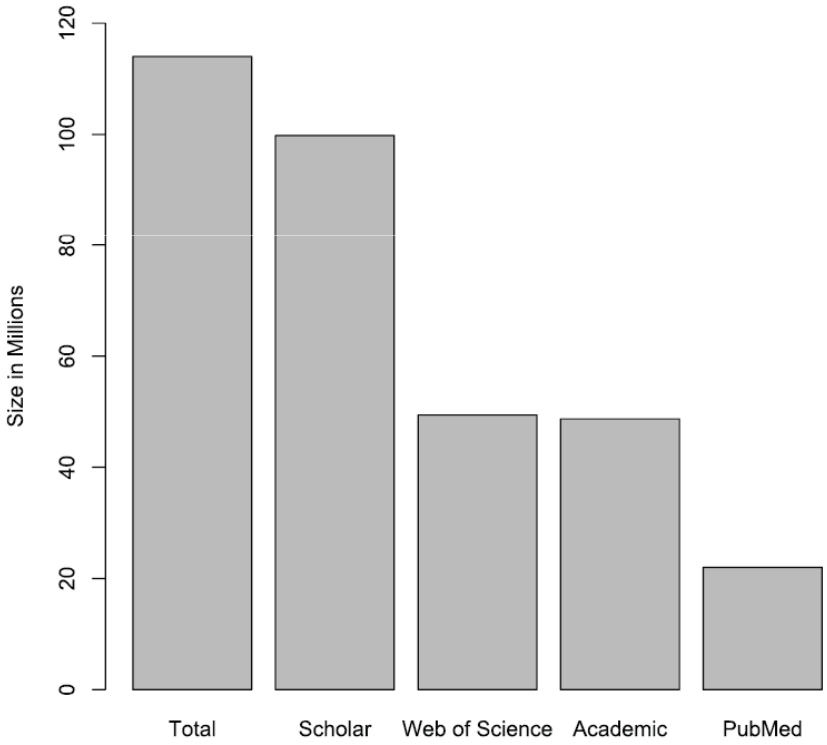
- the articles cited tend to be **more recent**
- there are **fewer citations**
- most citations are to **fewer journals and articles**
- **weakening ability to explore scientific literature laterally** finding in other studies and disciplines information potentially relevant to our current research

Source: STM Report 2015

Source: Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. science, 321(5887)

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:

Field	% of Public
Agriculture Science	12
Arts & Humanities	24
Biology	25
Chemistry	22
Computer Science	50
Economics & Business	42
Engineering	12
Environmental Sciences	29
Geosciences	35
Material Science	12
Mathematics	27
Medicine	26
Physics	35
Social Science	19
Multidisciplinary	43

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

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
share research

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

Scientific Literature Overload

Social Media in academic communication

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

Full paper notifications were already sent yesterday. In case you haven't received yours, please email ICDM2016Chairs@eurecat.org. Thanks.

#ICLR2017 Effective Quantization Methods for Recurrent Neural Networks. (arXiv:1611.10176v1 [cs.LG]) arxiv.org/abs/1611.10176

@neurobongo just read your paper arXiv:1511.06380. Great intuition!!!

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:

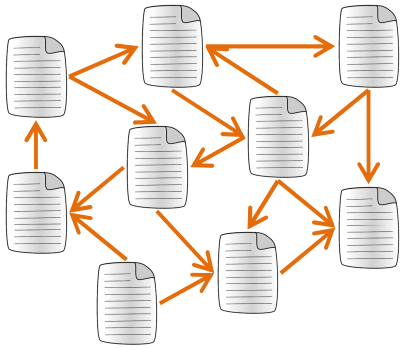
- **lack of clearly compelling benefits** with respect to the time needed to publishing material and manage interactions
- **quality** and **trust** issues

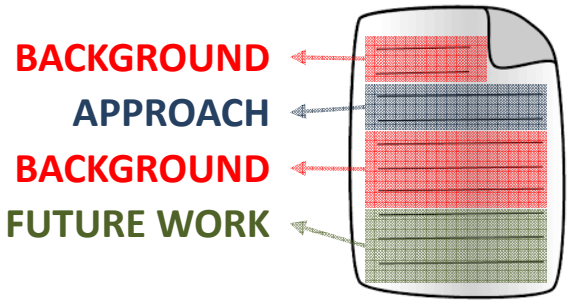
A wider adoption of altmetrics can foster the use of Social Media in scholarly communication

Sugimoto et al. (2016). Scholarly use of social media and altmetrics: a review of the literature.

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**





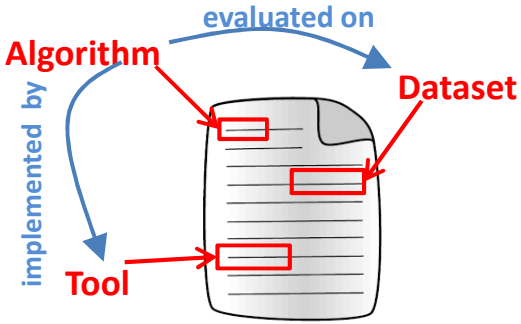
Citation network / data linking

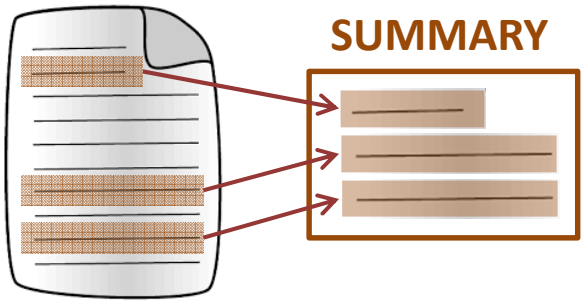
Scientific discourse

Entities and relations

Content relevance

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

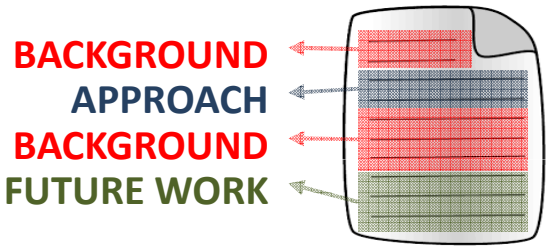




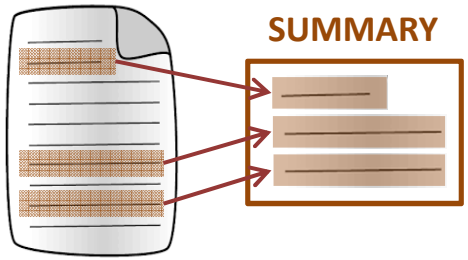
Text mining opportunities and challenges

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

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:**
http://europa.eu/rapid/press-release_MEMO-16-3011_en.htm
"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."*

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

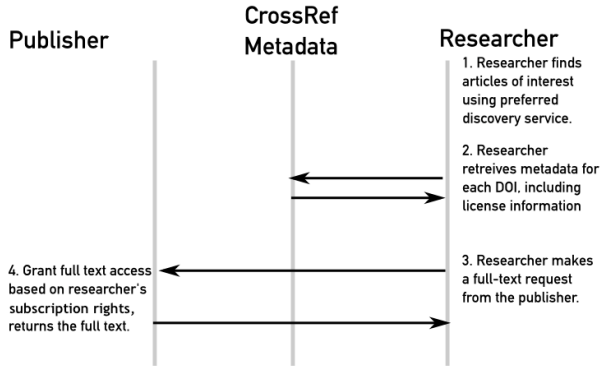
Crossref

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

Document Structure Analysis

Scientific Discourse Characterization

Citation Analysis

Scientific Document Summarization

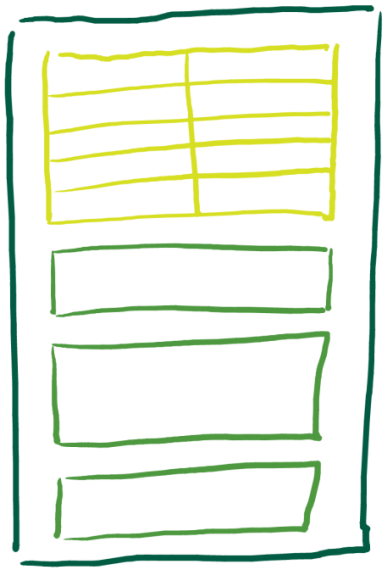
Universitat Pompeu Fabra Barcelona

taln

EXCELENCIA MARÍA DE MAEZTU

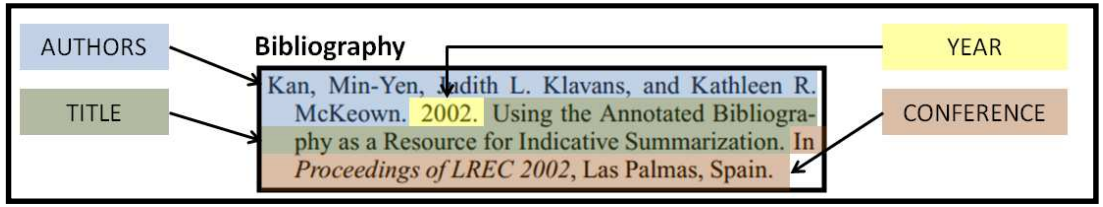
DOCUMENT STRUCTURE ANALYSIS

Publication



Header

Sections



Outline

- Document formats: dealing with PDF
- Genral-purpose PDF-to-text tools
- PDF-to-text for scientific publications
- Comparing PDF-to-text for scientific publications
- Bibliographic entry parsing
- Annotated datasets
- Conclusions

Document formats: dealing with PDF

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

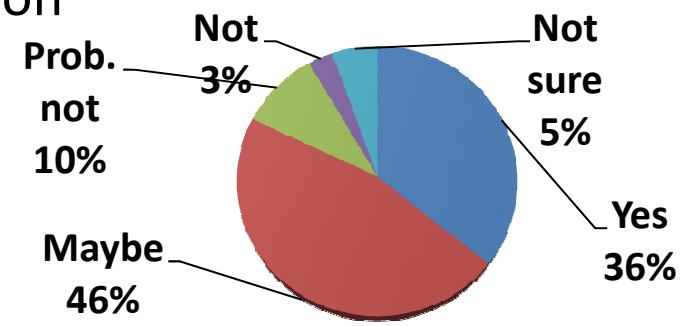
Document formats: dealing with PDF

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

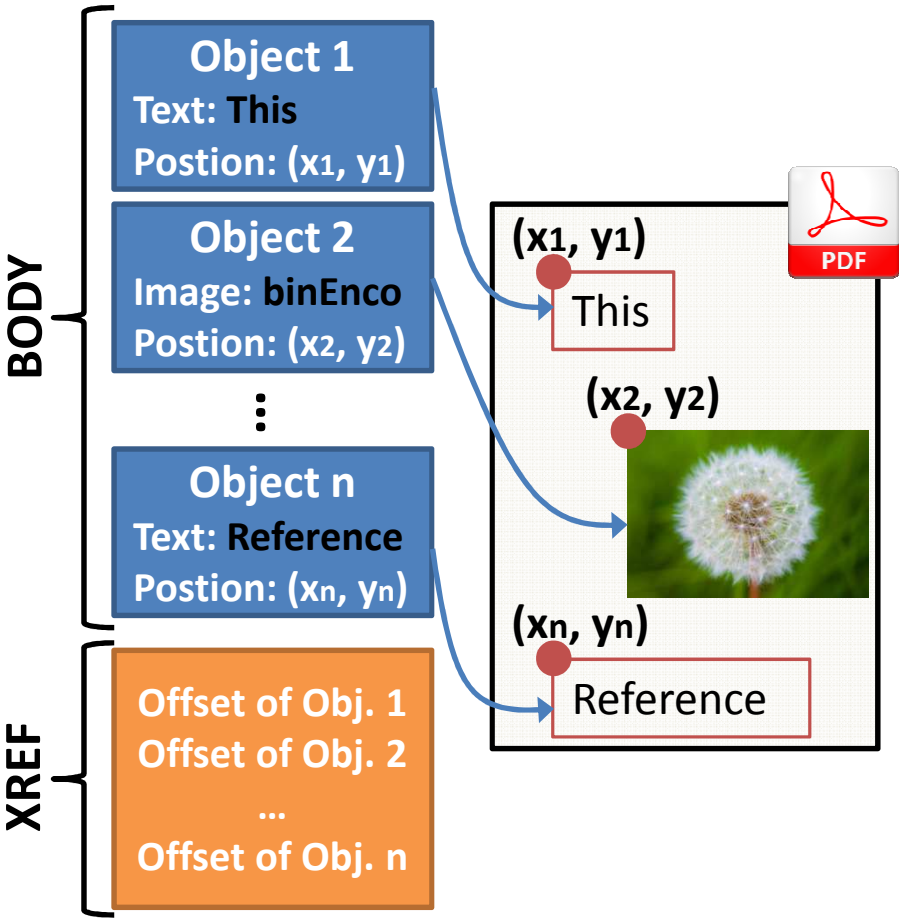
Do you expect that the way they access and use articles today to change in the future?
(281 responses of researchers in a variety of fields)



Document formats: dealing with PDF

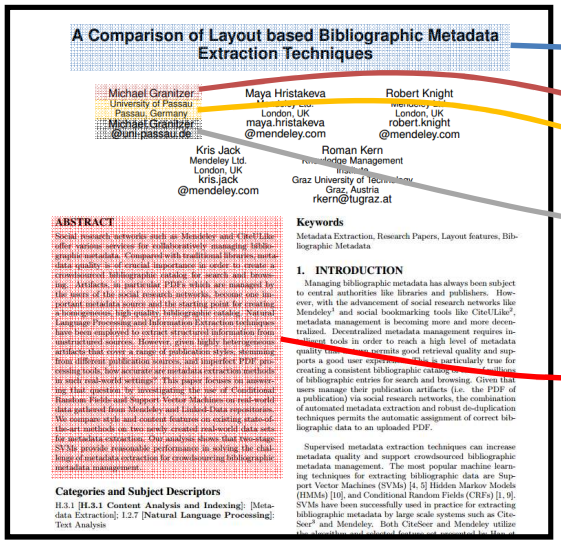
PDF is a layout based data format for professional document rendering

- 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)



Document formats: dealing with PDF

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

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 IDR solutions

Commercial
Specialized to extract tables, word lists and other elements

…also Adobe Acrobat , iText https://github.com/itext/itext7 , etc.

Try and compare these tools and others at: http://backingdata.org/pdfconv/

General purpose PDF-to-text software

Example of XML output generated by

pdf2xml

```
<DOCUMENT>
 <METADATA>
  <PDFFILENAME>/tmp/phpDKfZyT</PDFFILENAME>
  <PROCESS name="pdftoxml" cmd="-blocks -verbose ">
   <VERSION value="2.0">
    <COMMENT/>
   </VERSION>
   <CREATIONDATE>Mon Feb 17 20:22:55 2014</CREATIONDATE>
  </PROCESS>
 </METADATA>
 <PAGE width="595.276" height="841.89" number="1" id="p1">
  <MEDIABOX x1="0" y1="0" x2="595.276" y2="841.89"/>
  <CROPBOX x1="0" y1="0" x2="595.276" y2="841.89"/>
  <BLEEDBOX x1="0" y1="0" x2="595.276" y2="841.89"/>
  <ARTBOX x1="0" y1="0" x2="595.276" y2="841.89"/>
  <TRIMBOX x1="0" y1="0" x2="595.276" y2="841.89"/>
  <IMAGE id="p1_i1" sid="p1_s466" x="314.411" y="407.115" width="59.7845" height="93.3703"
  href="phpDKfZyT.pdftoxmlPDFTOXMLblocks.xml_data/image-1.png" clipZone="p1_c4"/>
<BLOCK id="p1_b1">
 <TEXT width="259.634" height="7.12527" x="72" y="60.4653" id="p1_t1">
  <TOKEN sid="p1_s3" id="p1_w1" angle-skewing-y="0" angle-skewing-x="0" leading="0" render="0" rise="0" horiz-scaling="1" word-space="0"
  char-space="0" font-name="NimbusRomNo9L" bold="no" italic="no" font-size="7.9701" font-color="#000000" rotation="0" angle="0" x="72"
  y="60.4653" base="65.869" width="60.788" height="7.12527">EUROGRAPHICS</TOKEN>
  <TOKEN sid="p1_s4" id="p1_w2" angle-skewing-y="0" angle-skewing-x="0" leading="0" render="0" rise="0" horiz-scaling="1" word-space="0"
  char-space="0" font-name="nimbusromno9l" bold="no" italic="no" font-size="7.9701" font-color="#000000" rotation="0" angle="0" x="134.78"
  y="60.4653" base="65.869" width="38.0891" height="7.12527">Symposium</TOKEN>
  <TOKEN sid="p1_s5" id="p1_w3" angle-skewing-y="0" angle-skewing-x="0" leading="0" render="0" rise="0" horiz-scaling="1" word-space="0"
  char-space="0" font-name="nimbusromno9l" bold="no" italic="no" font-size="7.9701" font-color="#000000" rotation="0" angle="0" x="174.862"
  y="60.4653" base="65.869" width="7.9701" height="7.12527">on</TOKEN>
  <TOKEN sid="p1_s6" id="p1_w4" angle-skewing-y="0" angle-skewing-x="0" leading="0" render="0" rise="0" horiz-scaling="1" word-space="0"
  char-space="0" font-name="nimbusromno9l" bold="no" italic="no" font-size="7.9701" font-color="#000000" rotation="0" angle="0" x="184.825"
  y="60.4653" base="65.869" width="43.7479" height="7.12527">Sketch-Based</TOKEN>
  <TOKEN sid="p1_s7" id="p1_w5" angle-skewing-y="0" angle-skewing-x="0" leading="0" render="0" rise="0" horiz-scaling="1" word-space="0"
  char-space="0" font-name="nimbusromno9l" bold="no" italic="no" font-size="7.9701" font-color="#000000" rotation="0" angle="0" x="230.565"
  y="60.4653" base="65.869" width="31.3384" height="7.12527">Interfaces</TOKEN>
  <TOKEN sid="p1_s8" id="p1_w6" angle-skewing-y="0" angle-skewing-x="0" leading="0" render="0" rise="0" horiz-scaling="1" word-space="0"
  char-space="0" font-name="nimbusromno9l" bold="no" italic="no" font-size="7.9701" font-color="#000000" rotation="0" angle="0" x="263.896"
  y="60.4653" base="65.869" width="11.5088" height="7.12527">and</TOKEN>
  <TOKEN sid="p1_s9" id="p1_w7" angle-skewing-y="0" angle-skewing-x="0" leading="0" render="0" rise="0" horiz-scaling="1" word-space="0"
  char-space="0" font-name="nimbusromno9l" bold="no" italic="no" font-size="7.9701" font-color="#000000" rotation="0" angle="0" x="277.397"
  y="60.4653" base="65.869" width="30.9957" height="7.12527">Modeling</TOKEN>
  <TOKEN sid="p1_s10" id="p1_w8" angle-skewing-y="0" angle-skewing-x="0" leading="0" render="0" rise="0" horiz-scaling="1" word-space="0"
  char-space="0" font-name="nimbusromno9l" bold="no" italic="no" font-size="7.9701" font-color="#000000" rotation="0" angle="0" x="310.386"
  y="60.4653" base="65.869" width="21.2483" height="7.12527">2011</TOKEN>
 </TEXT>
```

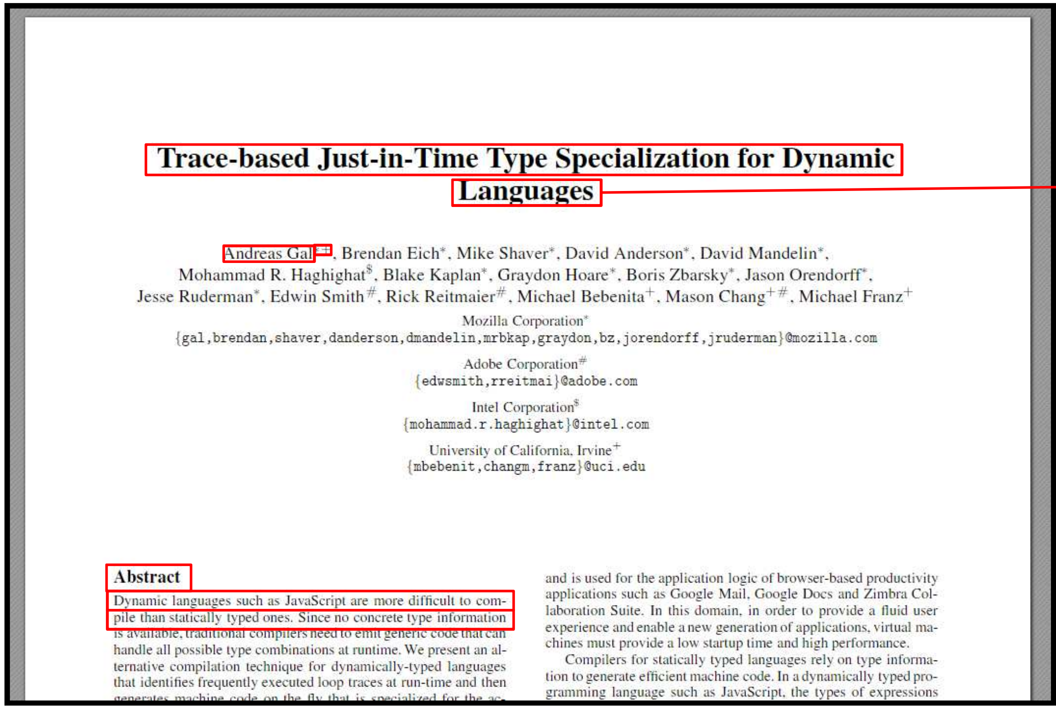
General purpose PDF-to-text software

Used by: readcube

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 Gal*+, Brendan Eich*, Mike Shaver*, David Anderson*, David Mandelin*, Mohammad R. Haghighat$, Blake Kaplan*, Graydon Hoare*, Boris Zbarsky*, Jason Orendorff*, Jesse Ruderman*, Edwin Smith#, Rick Reitmaier#, Michael Bebenita+, Mason Chang+#, Michael Franz+

Mozilla Corporation*
{gal,brendan,shaver,danderson,dmandelin,mrbkap,graydon,bz,jorendorff,jruderman}@mozilla.com

Adobe Corporation#
{edwsmith,rreitmai}@adobe.com

Intel Corporation$
{mohammad.r.haghighat}@intel.com

University of California, Irvine+
{mbebenit,changm,franz}@uci.edu

Abstract

Dynamic languages such as JavaScript are more difficult to compile than statically typed ones. Since no concrete type information is available, traditional compilers need to emit generic code that can handle all possible type combinations at runtime. We present an alternative compilation technique for dynamically-typed languages that identifies frequently executed loop traces at run-time and then generates machine code on the fly that is specialized for the ac-

and is used for the application logic of browser-based productivity applications such as Google Mail, Google Docs and Zimbra Collaboration Suite. In this domain, in order to provide a fluid user experience and enable a new generation of applications, virtual machines must provide a low startup time and high performance.

Compilers for statically typed languages rely on type information to generate efficient machine code. In a dynamically typed programming language such as JavaScript, the types of expressions

HTML source

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

CSS CLASSES TO DEFINE THE LAYOUT

```
.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**

CERMINE
Content ExtRactor and MINEr

GROBID

pdfx v1.9


lapdftext

SectLabel

MENDELEY

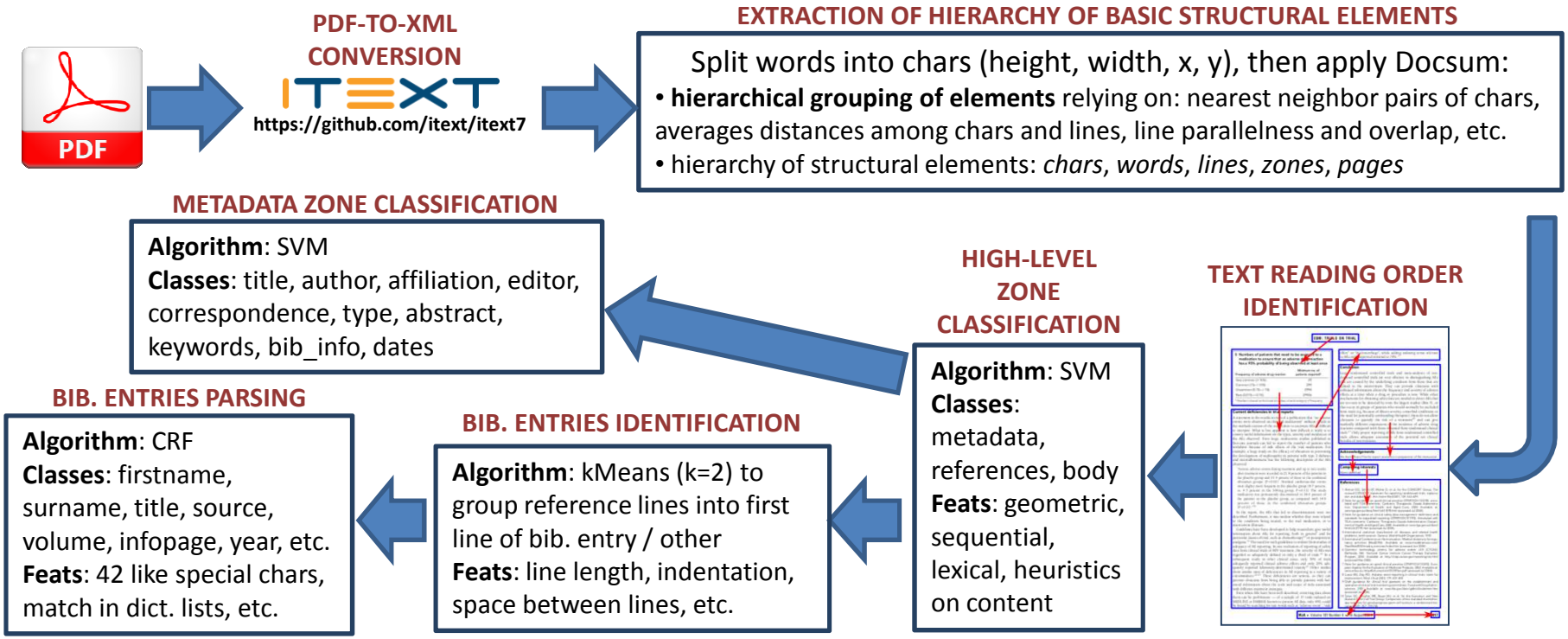
CiteSeer^{x} 10M

PDF-to-text for scientific publications

CERMINE
Content ExtRactor and MINEr

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

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



Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). *CERMINE: automatic extraction of structured metadata from scientific literature.* nternational Journal on Document Analysis and Recognition (IJDAR), 18(4), 317-335.

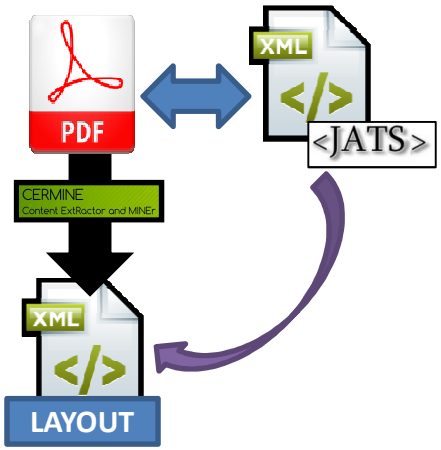
Document Structure Analysis

PDF-to-text for scientific publications

CERMINE
Content ExtRactor and MINEr

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

TRAINING AND EVALUAITON 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%

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). *CERMINE: automatic extraction of structured metadata from scientific literature.* nternational Journal on Document Analysis and Recognition (IJDAR), 18(4), 317-335.

PDF-to-text for scientific publications

GROBID

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

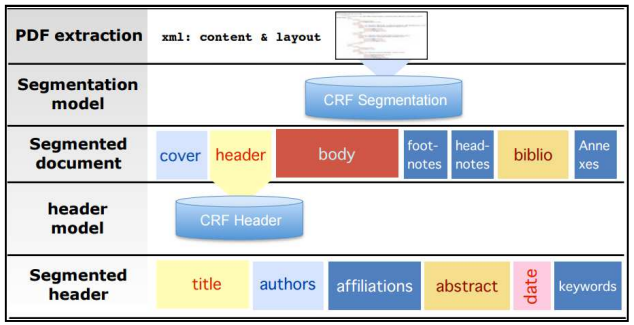
- Java (with JNI call to native CRF libraries: CRF++ or Wapiti) open-source (GitHub)
- PDF analysis with CRF exploiting both layout features and contents
- Output: TEI XML

PDF → PDF-TO-XML CONVERSION (token info) → pdf2xml

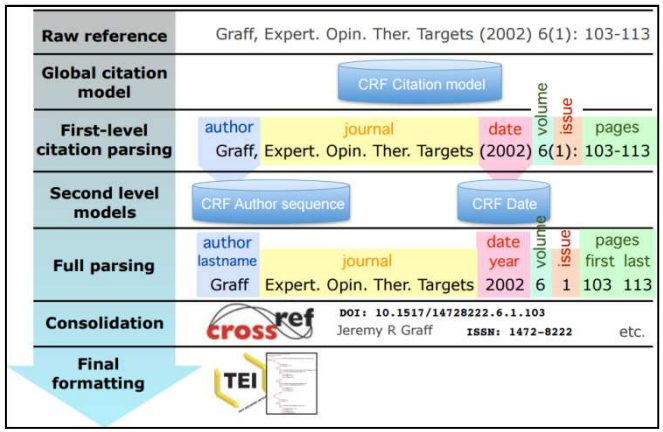
HIERARCHY OF CRF SEQUENCE TAGGERS

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

Document segmentation hierarchy



Bibliographic entry segmentation hierarchy



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

Lopez, P. (2009, September). *GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications.* In International Conference on Theory and Practice of Digital Libraries (pp. 473-474). Springer Berlin Heidelberg.

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)

Model	Nº training examples	Exploit layout
segmentation	121	x
header	3971	x
affiliation-address	1064	
names (header)	1297	
names (citation)	253	
date	619	
reference-segmenter	17	x
citation	4150	
fulltext (body)	8 (+13 abstracts)	x

A customized versions of GROBID is exploited by:

ResearchGate

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

Lopez, P. (2009, September). *GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications.* In International Conference on Theory and Practice of Digital Libraries (pp. 473-474). Springer Berlin Heidelberg.

PDF-to-text for scientific publications

pdfx v1.9

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 o the surrounding elements)

Front Matter	Body Matter	Back Matter / Others
title	body text	bibliographic item
author	(sub)section	(reference)
abstract	(sub)section heading	URI
author footnote	image	email
	table	side note
	caption	header/footer
	figure/table reference	page number
	bibliographic reference (citation)	

Constantin, A., Pettifer, S., & Voronkov, A. (2013, September). *PDFX: fully-automated PDF-to-XML conversion of scientific literature.* In Proceedings of the 2013 ACM symposium on Document engineering (pp. 177-180). ACM.

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

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

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

Constantin, A., Pettifer, S., & Voronkov, A. (2013, September). *PDFX: fully-automated PDF-to-XML conversion of scientific literature.* In Proceedings of the 2013 ACM symposium on Document engineering (pp. 177-180). ACM.

Document Structure Analysis

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

PDF-TO-XML CONVERSION (line based repr.)

Omnipage

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**

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

Luong, M. T., Nguyen, T. D., & Kan, M. Y. (2012). *Logical structure recovery in scholarly articles with rich document features.* Multimedia Storage and Retrieval Innovations for Digital Library Systems, 270.
Councill, I. G., Giles, C. L., & Kan, M. Y. (2008, May). *ParsCit: an Open-source CRF Reference String Parsing Package.* In LREC (Vol. 8, pp. 661-667).

PDF-to-text for scientific publications

https://github.com/BMKEG/lapdftextProject

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

PDF-TO-XML CONVERSION (word blocks detection)

JPedal

IDR solutions

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

```
rule "Title"
    activation-group "blockClassification"
    salience 4
    when
            ChunkFeatures(pageNumber==1)
             ChunkFeatures(mostPopularFontSize==20)

            eval(chunk.getNumberOfLine()<=6)
            ChunkFeatures(allignedMiddle==true)

    then
        chunk.setType(chunk.TYPE_TITLE);

end
```

Ramakrishnan, C., Patnia, A., Hovy, E., & Burns, G. A. (2012). *Layout-aware text extraction from full-text PDF of scientific articles.* Source code for biology and medicine, 7(1), 1.

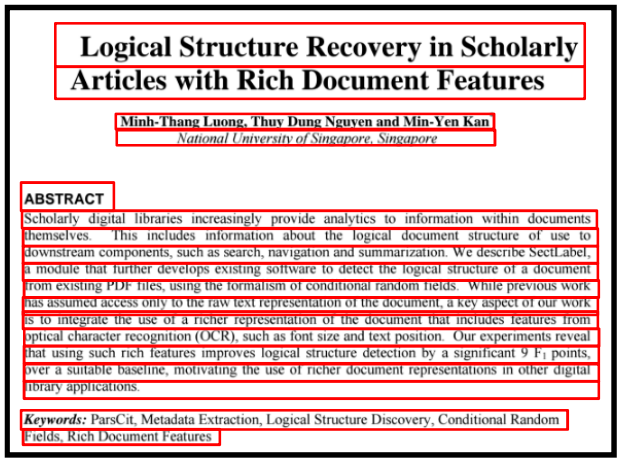
PDF-to-text for scientific publications

MENDELEY https://www.mendeley.com/download-mendeley-desktop/

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:

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2003*) Automatic document metadata extraction using support vector machines.* In Digital Libraries proceedings. 2003 Joint Conference IEEE.



Logical Structure Recovery in Scholarly Articles with Rich Document Features

Minh-Thang Luong, Thuy Dung Nguyen and Min-Yen Kan
National University of Singapore, Singapore

ABSTRACT

Scholarly digital libraries increasingly provide analytics to information within documents themselves. This includes information about the logical document structure of use to downstream components, such as search, navigation and summarization. We describe SectLabel, a module that further develops existing software to detect the logical structure of a document from existing PDF files, using the formalism of conditional random fields. While previous work has assumed access only to the raw text representation of the document, a key aspect of our work is to integrate the use of a richer representation of the document that includes features from optical character recognition (OCR), such as font size and text position. Our experiments reveal that using such rich features improves logical structure detection by a significant 9 F_1 points, over a suitable baseline, motivating the use of richer document representations in other digital library applications.

Keywords: ParsCit, Metadata Extraction, Logical Structure Discovery, Conditional Random Fields, Rich Document Features

1) **SVM that indipendently classifies each header line** with respect to textual features: position, number of words, number of capitalized words, % of words in specific dictionares, % of words occurring in a specific class of traning 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**

PDF-to-text for scientific publications

MENDELEY https://www.mendeley.com/download-mendeley-desktop/

How header metadata are extracted from a PDF document?

Later approach:

GROBID

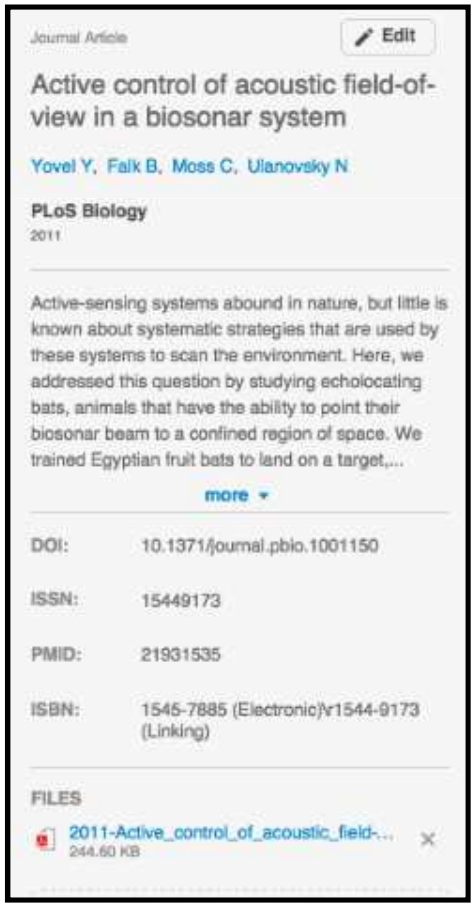
- Trained on a large set of papers
- **Fields**: title, authors, DOI, publication, volume, issue, year, page, ranges

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/



Document Structure Analysis

Comparing PDF-to-text for scientific publications

Lipinski, M., Yao, K., Breitinger, C., Beel, J., & Gipp, B. (2013, July).
Evaluation of header metadata extraction approaches and tools for scientific PDF documents.
In Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries (pp. 385-386). ACM.

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

	Title			Authors			Authors' last names		Abstract			Year	
	A_{100}	B_{100}	B_{1153}	A_{100}	B_{100}	B_{1153}	B_{100}	B_{1153}	A_{100}	B_{100}	B_{1153}	B_{100}	B_{1153}
GROBID	N/A	0.92	0.92	N/A	0.83	0.83	0.90	0.91	N/A	0.75	0.74	0.64	0.69
Mendeley Desktop	N/A	0.84	0.82	N/A	0.72	0.70	0.78	0.77	N/A	N/A	N/A	0.23	0.26
ParsCit SectLabel	0.59	0.52	0.54	0.47	0.29	0.31	0.36	0.37	0.49	0.31	0.26	0.06	0.07
PDFSSA4MET	0.13	0.21	0.18	0.05	0.02	0.01	0.20	0.18	N/A	N/A	N/A	N/A	N/A
PDFMeat	0.60	N/A	N/A	0.6	N/A	N/A	N/A	N/A	0.14	N/A	N/A	N/A	N/A
SciPlore Xtract	0.76	0.81	0.78	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SVMHeaderParse	0.50	0.57	0.61	0.64	0.70	0.73	0.74	0.76	0.37	0.64	0.64	0.21	0.20

Accuracy values

Document Structure Analysis

Comparing PDF-to-text for scientific publications

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015).
CERMINE: automatic extraction of structured metadata from scientific literature.
nternational Journal on Document Analysis and Recognition (IJDAR), 18(4), 317-335.

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)

	CERMINE	PDFX	GROBID	ParsCit	Pdf-extract
Title	**95.5**	85.7	82.5	34.1	49.4
	93.4	84.7	77.4	39.6	49.4
	94.5	85.2	79.8	36.6	49.4
Authors	**90.2**	71.2	85.9	57.9	–
	89.0	71.5	**90.5**	48.6	–
	89.6	71.3	88.1	52.8	–
Affiliations	88.2	–	**90.8**	72.2	–
	83.1	–	51.8	44.3	–
	85.6	–	66.0	54.9	–
Email addresses	51.7	**53.0**	26.9	28.8	–
	42.6	**73.6**	7.8	36.2	–
	46.7	**61.6**	12.1	32.1	–
Abstract	**82.8**	71.1	70.4	47.7	–
	79.9	66.7	67.7	61.3	–
	81.3	68.8	69.0	53.7	–
Keywords	89.9	–	**94.2**	15.6	–
	63.5	–	44.2	3.0	–
	74.4	–	60.2	5.1	–
Journal	**80.3**	–	–	–	–
	73.2	–	–	–	–
	76.6	–	–	–	–

	CERMINE	PDFX	GROBID	ParsCit	Pdf-extract
Volume	**93.3**	–	–	–	–
	83.0	–	–	–	–
	87.8	–	–	–	–
Issue	**53.7**	–	–	–	–
	28.4	–	–	–	–
	37.1	–	–	–	–
Pages	**87.0**	–	–	–	–
	80.4	–	–	–	–
	83.5	–	–	–	–
Year	**96.3**	–	95.7	–	–
	95.0	–	40.4	–	–
	95.6	–	56.8	–	–
DOI	98.2	–	**99.1**	–	–
	75.0	–	65.4	–	–
	85.1	–	78.8	–	–
References	**96.1**	91.3	79.7	81.2	80.4
	89.8	88.9	66.7	71.8	57.5
	92.8	90.1	72.6	76.2	67.0

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/
  - **ParsCit**: CRF++ library, open source (perl and C++)
    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
  - **Bibsonomy REST API – search posts by string**:
    https://bitbucket.org/bibsonomy/bibsonomy/wiki/documentation/api/REST%20API
    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

Seymore, K., McCallum, A., & Rosenfeld, R. (1999, July). *Learning hidden Markov model structure for information extraction.* In AAAI-99 Workshop on Machine Learning for Information Extraction (pp. 37-42).

- 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

Cortez, E., da Silva, A. S., Gonçalves, M. A., Mesquita, F., & de Moura, E. S. (2007, June). *FLUX-CIM: flexible unsupervised extraction of citation metadata.* In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries(pp. 215-224). ACM.

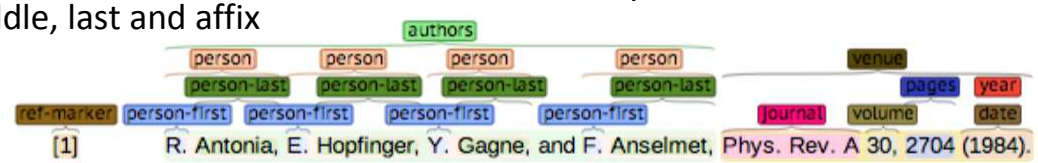
- 300 citation strings randomly from ACM Digital Library - CORA format

UMASS Citation dataset

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

Anzaroot, S., & McCallum, A. (2013). *A new dataset for fine-grained citation field extraction.* In ICML Workshop on Peer Reviewing and Publishing Models.

- 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



Annotated datasets

CiteSeer Citation dataset

https://github.com/knmnyn/ParsCit/blob/master/doc/citeseerx.tagged.txt

Lawrence, S., Giles, C. L., & Bollacker, K. D. (1999, April*). Autonomous citation matching.* In Proceedings of the third annual conference on Autonomous Agents (pp. 392-393). ACM.

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

GROTOAP2 - GROund Truth for Open Access Publications

http://cermine.ceon.pl/grotoap2/

Tkaczyk, D., Szostek, P., & Bolikowski, L. (2014). GROTOAP2 The Methodology of Creating a Large Ground Truth Dataset of Scientific Articles. D-Lib Magazine, 20(11), 13.

- **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, aknowledgments, 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

Universitat Pompeu Fabra Barcelona

EXCELENCIA MARÍA DE MAEZTU

SCIENTIFIC DISCOURSE CHARACTERIZATION

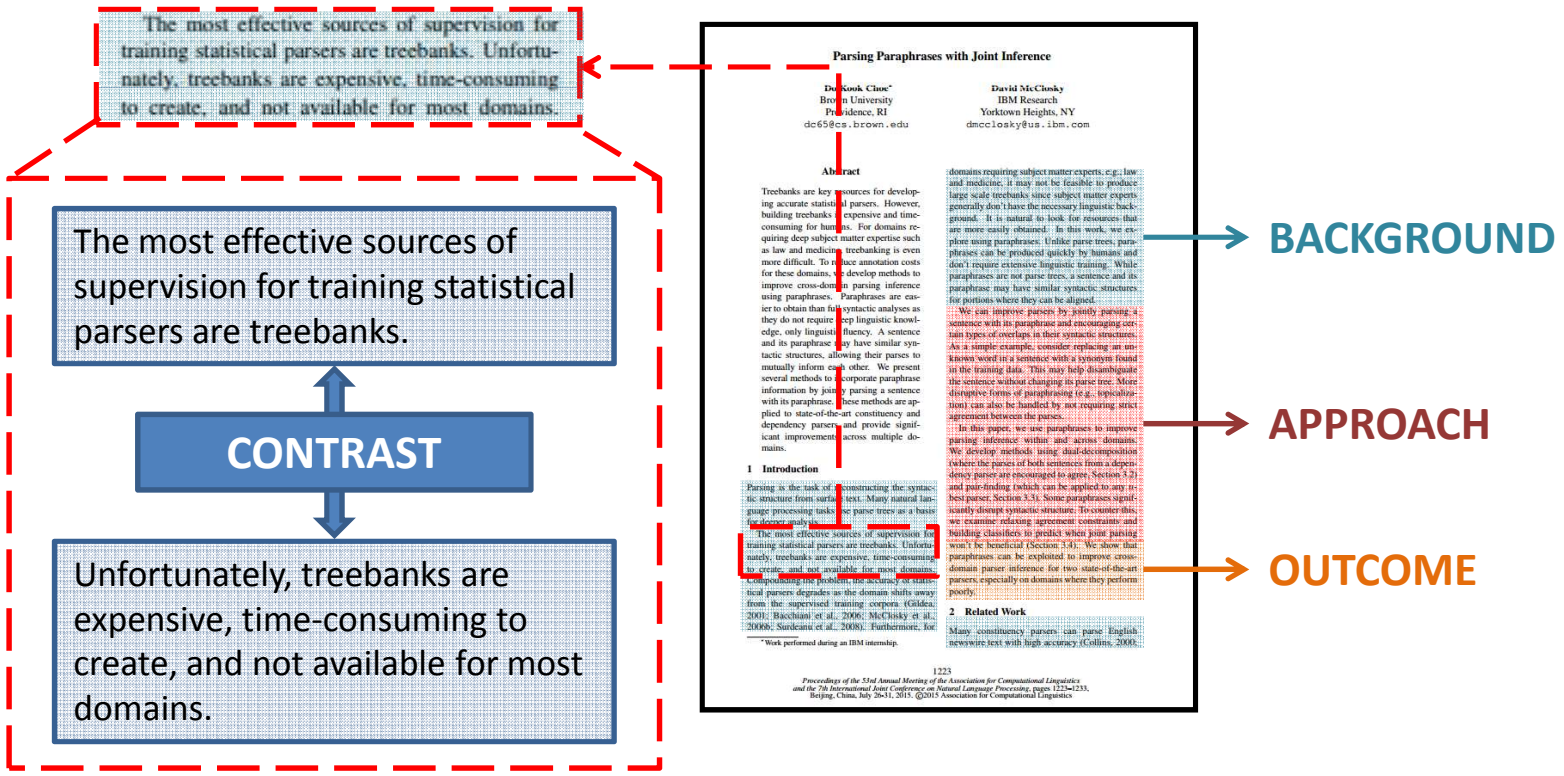
Outline

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

Scientific Discourse Characterization

What is scientific discourse?

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



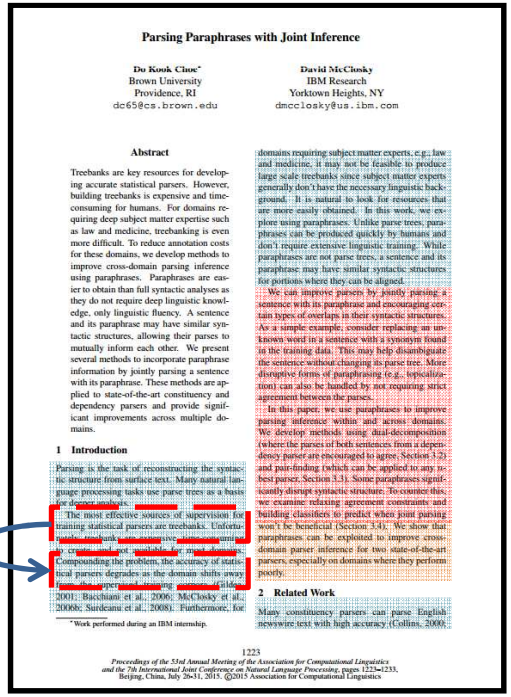
What is scientific discourse?

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**



Parsing Paraphrases with Joint Inference

Do Kook Choe*
Brown University
Providence, RI
dc65@cs.brown.edu

David McClosky
IBM Research
Yorktown Heights, NY
dmcclosky@us.ibm.com

Abstract

Treebanks are key resources for developing accurate statistical parsers. However, building treebanks is expensive and time-consuming for humans. For domains requiring deep subject matter expertise such as law and medicine, treebanking is even more difficult. To reduce annotation costs for these domains, we develop methods to improve cross-domain parsing inference using paraphrases. Paraphrases are easier to obtain than full syntactic analyses as they do not require deep linguistic knowledge, only linguistic fluency. A sentence and its paraphrase may have similar syntactic structures, allowing their parses to mutually inform each other. We present several methods to incorporate paraphrase information by jointly parsing a sentence with its paraphrase. These methods are applied to state-of-the-art constituency and dependency parsers and provide significant improvements across multiple domains.

1 Introduction

Parsing is the task of reconstructing the syntactic structure from surface text. Many natural language processing tasks use parse trees as a basis for deeper analysis.

The most effective sources of supervision for training statistical parsers are treebanks. Unfortunately, treebanks are expensive, time-consuming to create, and not available for many domains. Compounding the problem, the accuracy of statistical parsers degrades as the domain shifts away from the parser's training [illegible] (Gildea, 2001; Bacchiani et al., 2006; McClosky et al., 2006b; Surdeanu et al., 2008). Furthermore, for domains requiring subject matter experts, e.g. law and medicine, it may not be feasible to produce large scale treebanks since subject matter experts generally don't have the necessary linguistic background. It is natural to look for resources that are more easily obtained. In this work, we explore using paraphrases. Unlike parse trees, paraphrases can be produced quickly by humans and don't require extensive linguistic training. While paraphrases are not parse trees, a sentence and its paraphrase may have similar syntactic structures for portions where they can be aligned.

We can improve parsers by jointly parsing a sentence with its paraphrase and encouraging certain types of overlaps in their syntactic structures. As a simple example, consider replacing an unknown word in a sentence with a synonym found in the training data. This may help disambiguate the sentence without changing its parse tree. More disruptive forms of paraphrasing (e.g., topicalization) can also be handled by not requiring strict agreement between the parses.

In this paper, we use paraphrases to improve parsing inference within and across domains. We develop methods using dual-decomposition (where the parses of both sentences from a dependency parser are encouraged to agree, Section 3.2) and pair-finding (which can be applied to any k-best parser, Section 3.3). Some paraphrases significantly disrupt syntactic structure. To counter this, we examine relaxing agreement constraints and building classifiers to predict when joint parsing won't be beneficial (Section 3.4). We show that paraphrases can be exploited to improve cross-domain parser inference for two state-of-the-art parsers, especially on domains where they perform poorly.

2 Related Work

Many constituency parsers can parse English newswire text with high accuracy (Collins, 2000;

* Work performed during an IBM internship.

1223

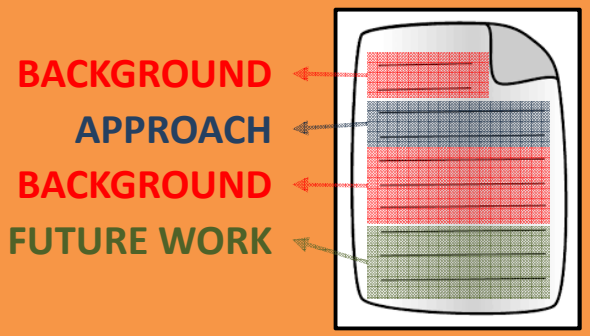
Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1223–1233, Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

Scientific discourse characterization

Steps towards the characterization and automated annotation of scientific discourse

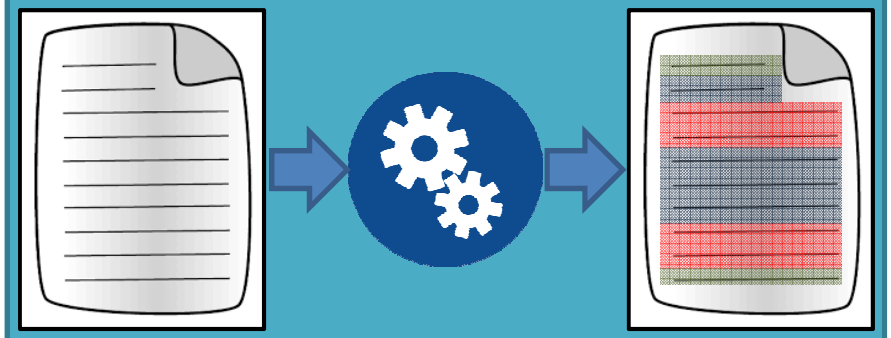
Annotation procedure and annotated corpus

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



Automated annotation of scientific texts

- algorithmic approach
- feature engineering



Scientific discourse characterization

Annotation procedure and annotated corpus

Two main approaches to the rhetorical analysis of a text:

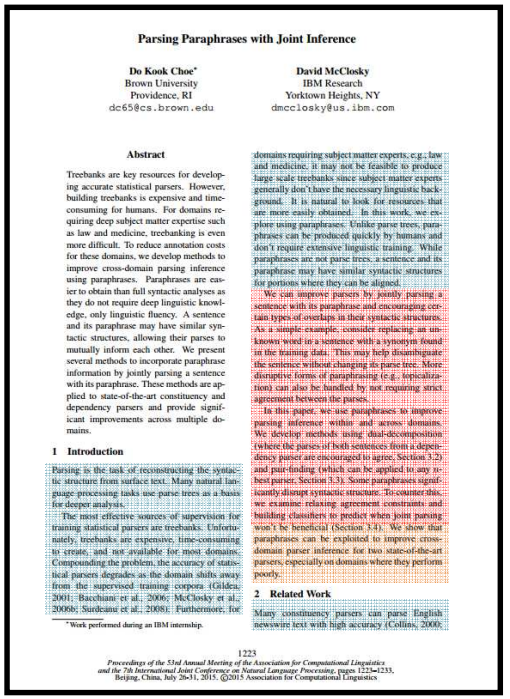
Rhetorical Structure Theory: relations between clauses or larger text segments

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.

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



Parsing Paraphrases with Joint Inference

Do Kook Choe* — Brown University, Providence, RI — dc65@cs.brown.edu

David McClosky — IBM Research, Yorktown Heights, NY — dmcclosky@us.ibm.com

Abstract

Treebanks are key resources for developing accurate statistical parsers. However, building treebanks is expensive and time-consuming for humans. For domains requiring deep subject matter expertise such as law and medicine, treebanking is even more difficult. To reduce annotation costs for these domains, we develop methods to improve cross-domain parsing inference using paraphrases. Paraphrases are easier to obtain than full syntactic analyses as they do not require deep linguistic knowledge, only linguistic fluency. A sentence and its paraphrase may have similar syntactic structures, allowing their parses to mutually inform each other. We present several methods to incorporate paraphrase information by jointly parsing a sentence with its paraphrase. These methods are applied to state-of-the-art constituency and dependency parsers and provide significant improvements across multiple domains.

1 Introduction

Parsing is the task of reconstructing the syntactic structure from surface text. Many natural language processing tasks use parse trees as a basis for deeper analysis.

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. Compounding the problem, the accuracy of statistical parsers degrades as the domain shifts away from the supervised training corpora (Gildea, 2001; Bacchiani et al., 2006; McClosky et al., 2006b; Surdeanu et al., 2008). Furthermore, for domains requiring subject matter experts, e.g. law and medicine, it may not be feasible to produce large scale treebanks since subject matter experts generally don't have the necessary linguistic background. It is natural to look for resources that are more easily obtained. In this work, we explore using paraphrases. Unlike parse trees, paraphrases can be produced quickly by humans and don't require extensive linguistic training. While paraphrases are not parse trees, a sentence and its paraphrase may have similar syntactic structures for portions where they can be aligned.

We can improve parsers by jointly parsing a sentence with its paraphrase and encouraging certain types of overlaps in their syntactic structures. As a simple example, consider replacing an unknown word in a sentence with a synonym found in the training data. This may help disambiguate the sentence without changing its parse tree. More disruptive forms of paraphrasing (e.g., topicalization) can also be handled by not requiring strict agreement between the parses.

In this paper, we use paraphrases to improve parsing inference within and across domains. We develop methods using dual-decomposition (where the parses of both sentences from a dependency parser are encouraged to agree, Section 3.2) and pair-finding (which can be applied to any n-best parser, Section 3.3). Some paraphrases significantly disrupt syntactic structure. To counter this, we examine relaxing agreement constraints and building classifiers to predict when joint parsing won't be beneficial (Section 3.4). We show that paraphrases can be exploited to improve cross-domain parser inference for two state-of-the-art parsers, especially on domains where they perform poorly.

2 Related Work

Many constituency parsers can parse English newswire text with high accuracy (Collins, 2000;

* Work performed during an IBM internship.

1223

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1223–1233, Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

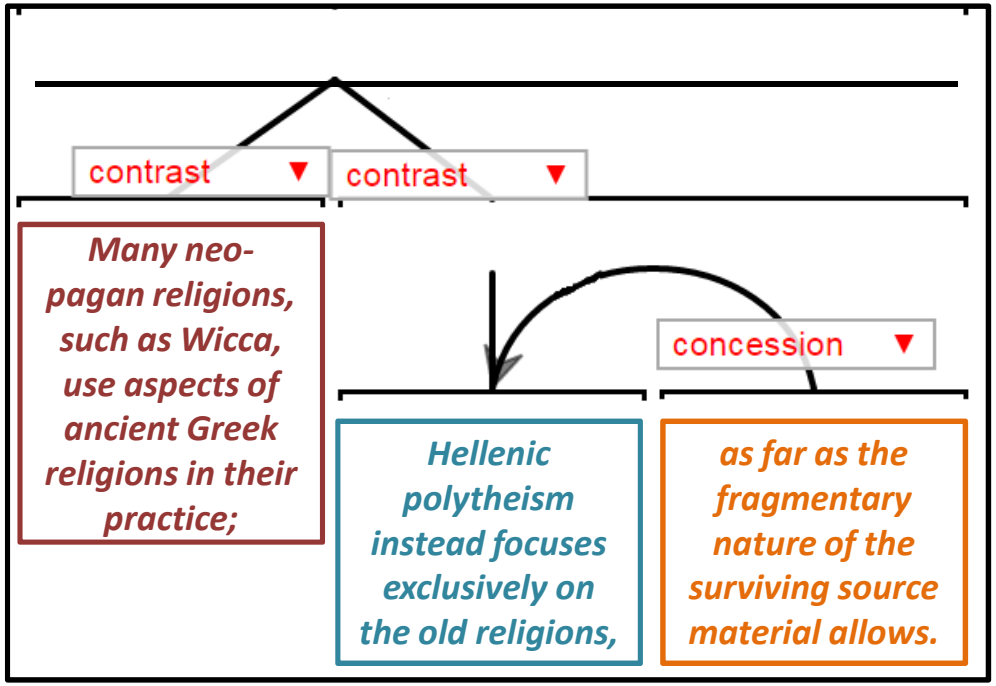
Scientific discourse characterization

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

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

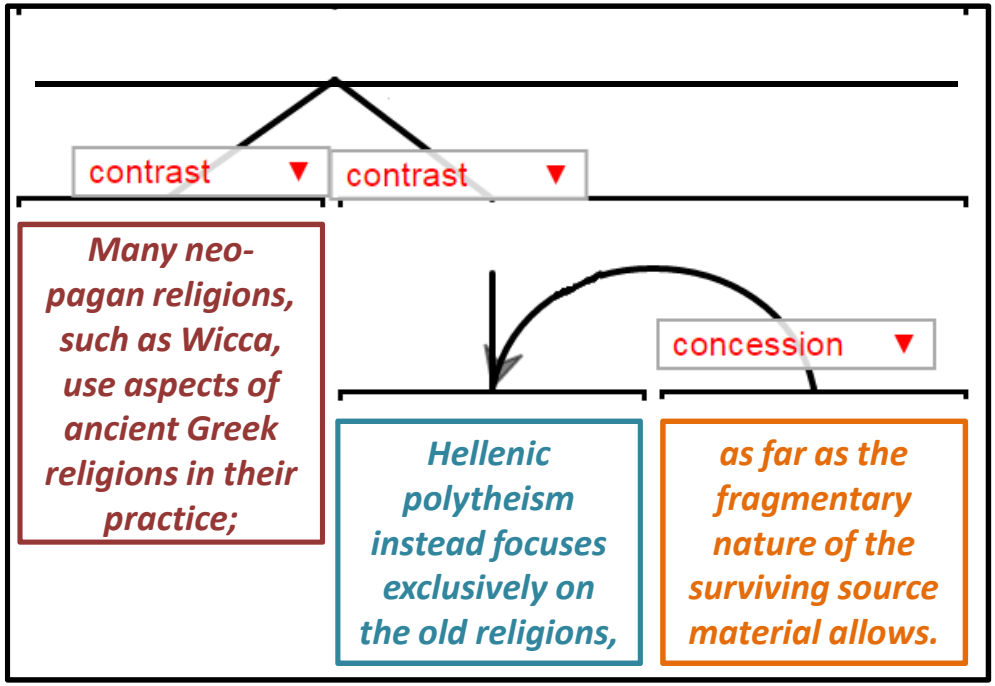
Annotation procedure and annotated corpus

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



Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis

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

Dr. Inventor Scientific Discourse Schema

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: Knowledge Claim Discourse Model

The argumentative structure of a scientific article is based on the need of authors to convince the reader of their contributions by **claiming the ownership of a new piece of knowledge**

*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***

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

EXAMPLES OF RHETORICAL MOVES

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.

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: Argumentative Zoning

Argumentative Zoning Annotation Schema bundles together similar rhetorical moves casting the general argumentation recognition Knowledge Claim Discourse Model into a sentence classification task

AZ Annotation Schema

Category	Description	
AIM	Statement of research goal.	2%
BACKGROUND	Description of generally accepted background knowledge.	6%
BASIS	Existing KC provides basis for new KC.	2%
CONTRAST	An existing KC is contrasted, compared, or presented as weak.	5%
OTHER	Description of existing KC.	16%
OWN	Description of any other aspect of new KC.	67%
TEXTUAL	Indication of paper's textual structure.	2%

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

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)

Category	Description
AIM	Statement of specific research goal, or hypothesis of current paper AZ Aim
NOV_ADV	Novelty or advantage of own approach
CO_GRO	No knowledge claim is raised (or knowledge claim not significant for the paper) AZ Background
OTHR	Knowledge claim (significant for paper) held by somebody else. Neutral description AZ Other
PREV_OWN	Knowledge claim (significant) held by authors in a previous paper. Neutral description.
OWN_MTHD	New Knowledge claim, own work: methods
OWN_FAIL	A solution/method/experiment in the paper that did not work AZ Own
OWN_RES	Measurable/objective outcome of own work

Category	Description
OWN_CONC	Findings, conclusions (non-measurable) of own work AZ Own
CODI	Comparison, contrast, difference to other solution (neutral)
GAP_WEAK	Lack of solution in field, problem with other solutions AZ Contrast
ANTISUPP	Clash with somebody else's results or theory; superiority of own work
SUPPORT	Other work supports current work or is supported by current work AZ Basis
USE	Other work is used in own work
FUT	Statements/suggestions about future work (own or general)

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

Annotation:
3 annotators experts in CL with different levels of expertise in Chemistry → chemistry intro

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)

Category	Description
AIM	Statement of specific research goal, or hypothesis of current paper
NOV_ADV	Novelty or advantage of own approach
CO_GRO	No knowledge claim is raised (or knowledge claim not significant for the paper)
OTHR	Knowledge claim (significant for paper) held by somebody else. Neutral description
PREV_OWN	Knowledge claim (significant) held by authors in a previous paper. Neutral description.
OWN_MTHD	New Knowledge claim, own work: methods
OWN_FAIL	A solution/method/experiment in the paper that did not work
OWN_RES	Measurable/objective outcome of own work

Category	Description
OWN_CONC	Findings, conclusions (non-measurable) of own work
CODI	Comparison, contrast, difference to other solution (neutral)
GAP_WEAK	Lack of solution in field, problem with other solutions
ANTISUPP	Clash with somebody else's results or theory; superiority of own work
SUPPORT	Other work supports current work or is supported by current work
USE	Other work is used in own work
FUT	Statements/suggestions about future work (own or general)

AZ Aim
AZ Background
AZ Other
AZ Own
AZ Own
AZ Contrast
AZ Basis

Inter-annotator agreement is comparable across domains (k is 0.65 in CL and 0.71 in Chemistry)

Higher agreement among Chemistry experts → a little improvement of annotation quality with domain knowledge

Scientific discourse characterization

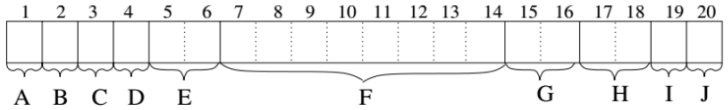
Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, *28*(4), 409-445.

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)

	AIM			CONTR.			TEXTUAL			OWN			BACKG.			BASIS			OTHER		
	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
System	52	44	65	26	34	20	61	57	66	86	84	88	45	40	50	38	37	40	44	52	39

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Teufel, S., & Kan, M. Y. (2011). *Robust argumentative zoning for sensemaking in scholarly documents* (pp. 154-170). Springer Berlin Heidelberg.

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

Teufel, S., & Kan, M. Y. (2011). *Robust argumentative zoning for sensemaking in scholarly documents* (pp. 154-170). Springer Berlin Heidelberg.

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

n=12898	# of instances	F_1	F-score with structural features
AIM	229 (1.77%)	51%	52%
BAS	155 (1.20%)	22%	38%
BKG	493 (3.82%)	24%	45%
CTR	302 (2.34%)	19%	26%
OTH	1598 (12.38%)	31%	44%
OWN	9889 (76.67%)	81%	86%
TXT	158 (1.22%)	61%	61%
Un.	74 (0.5%)	–	

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

Séaghdha, D. O., & Teufel, S. (2014). *Unsupervised learning of rhetorical structure with un-topic models.* In *COLING* (pp. 2-13).

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

Séaghdha, D. O., & Teufel, S. (2014). *Unsupervised learning of rhetorical structure with un-topic models.* In *COLING* (pp. 2-13).

Two language models can be composed by a binary-valued latent variable to generates 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.

Topic-independent template for abstracts of NLP papers, filled

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Séaghdha, D. O., & Teufel, S. (2014). *Unsupervised learning of rhetorical structure with un-topic models.* In *COLING* (pp. 2-13).

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 → most likely topic
- a distribution of rhetorical zones → 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.

→ **TOPIC N. 20 (over 100)**
ZONE N. 3 (over 10)

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Séaghdha, D. O., & Teufel, S. (2014). *Unsupervised learning of rhetorical structure with un-topic models.* In *COLING* (pp. 2-13).

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

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Séaghdha, D. O., & Teufel, S. (2014). *Unsupervised learning of rhetorical structure with un-topic models.* In *COLING* (pp. 2-13).

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

Merity, S., Murphy, T., & Curran, J. R. (2009, August). *Accurate argumentative zoning with maximum entropy models.* In Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (pp. 19-26). Association for Computational Linguistics.

- 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

Feltrim, V. D., Teufel, S., das Nunes, M. G. V., & Aluísio, S. M. (2006). *Argumentative zoning applied to critiquing novices' scientific abstracts.* In Computing Attitude and Affect in Text: Theory and Applications (pp. 233-246). Springer Netherlands.

- **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

Automated annotation of scientific texts

Zone Analysis: Argumentative Zoning

Mizuta, Y., & Collier, N. (2004, May). *An Annotation Scheme for a Rhetorical Analysis of Biology Articles.* In *LREC* (pp. 1737-1740).

- 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

Hachey, B., & Grover, C. (2006). *Extractive summarisation of legal texts. Artificial Intelligence and Law*, *14*(4), 305-345.

- 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

Annotation procedure and annotated corpus

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

Category	Description	
Hypothesis	An unconfirmed statemen which is a stepping stone of the investigation	2%
Motivation	The reason behind the investigation	1%
Background	Generally expected background knowledge and previous work	19%
Goal	A target state of the investigation where intended discoveries are made	1%
Object	An entity which is the product or main theme of the investigation (**advantage / disadvantage**)	3%
Method	Means by which the authors seek to achieve the goal of the investigation (**old / new – advantage / disadvantage**)	11%
Experiment	An experimental method	10%
Model	A statement about a theoretical model or framework	9%
Observation	The data / phenomena recorded in an investigation	14%
Result	Factual statements about the outputs, interpretation of an observation	21%
Conclusion	Statements inferred from observations and results	9%

ART Corpus

- **265 papers** from the domains of **chemistry and biochemistry**
- **39,915 sentences**
- Avg. annotator agreement K: **0,55**

Online at (SciXML format): https://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/

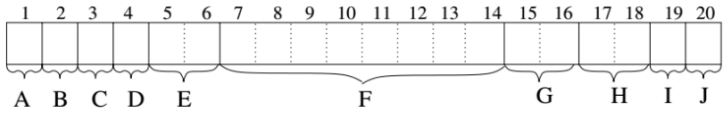
Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Core Scientific Concepts

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). *Automatic recognition of conceptualization zones in scientific articles and two life science applications.* Bioinformatics, 28(7)

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

Annotation procedure and annotated corpus

Zone Analysis: Core Scientific Concepts, multi-class

James Ravenscroft, Maria Liakata, Anika Oellrich, and Shyamasree Saha (2016). *Multi-label annotation in scientific articles – The Multi-label Cancer Risk Assessment Corpus.* LREC

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

Bone marrow stromal cells were treated with AhR agonists and bacterial → **METHOD**
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

Annotation procedure and annotated corpus

Zone Analysis: Core Scientific Concepts, multi-class

James Ravenscroft, Maria Liakata, Anika Oellrich, and Shyamasree Saha (2016). *Multi-label annotation in scientific articles – The Multi-label Cancer Risk Assessment Corpus.* LREC

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, tested on CRA corpus

New: trained and tested on CRA corpus

Label	F-measure	
	Old	New
Bac	52.6	**76.7**
Con	41.3	**53.9**
Exp	80.9	**83.4**
Goa	43.3	**54.2**
Hyp	11.6	**13.9**
Met	48.2	**58.3**
Mod	00.0	0.00
Mot	07.4	**43.9**
Obj	26.1	**27.0**
Obs	35.1	**43.0**
Res	47.5	**68.3**

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

Annotation procedure and annotated corpus

Zone Analysis: Argumentative Zoning vs Core Scientific Concepts

Liakata, M., Teufel, S., Siddharthan, A., & Batchelor, C. R. (2010, May). *Corpora for the Conceptualisation and Zoning of Scientific Papers.* In LREC.

AZ-II

characterize the **ownership of the knowledge claims** presented in the paper, thus identifying and motivating the new contributions of the author

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

CoreSC

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

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

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: Argumentative Zoning vs Core Scientific Concepts

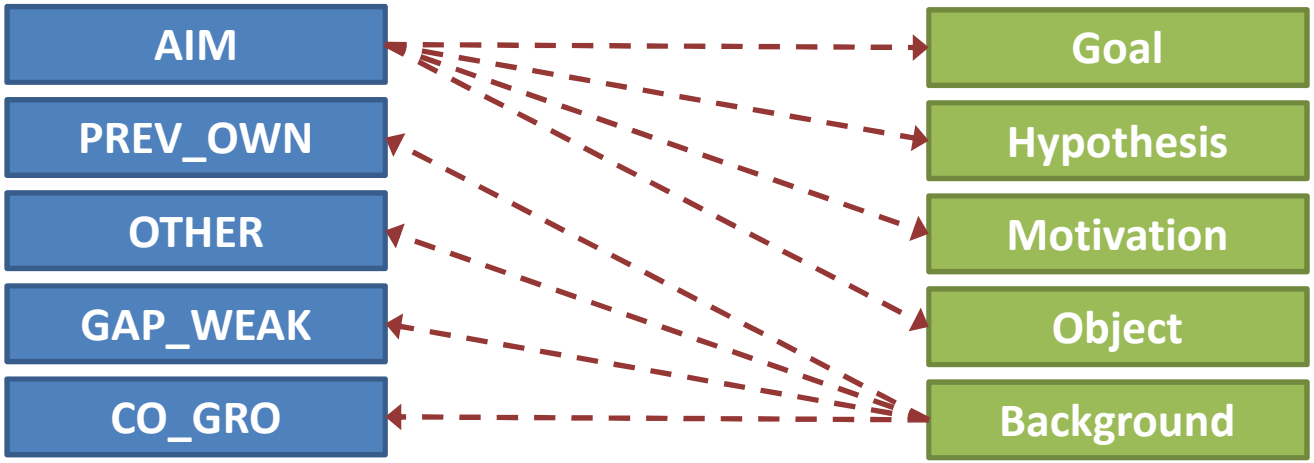
Liakata, M., Teufel, S., Siddharthan, A., & Batchelor, C. R. (2010, May). *Corpora for the Conceptualisation and Zoning of Scientific Papers.* In LREC.

AZ-II **CoreSC**

36 papers annotated with both schemas

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

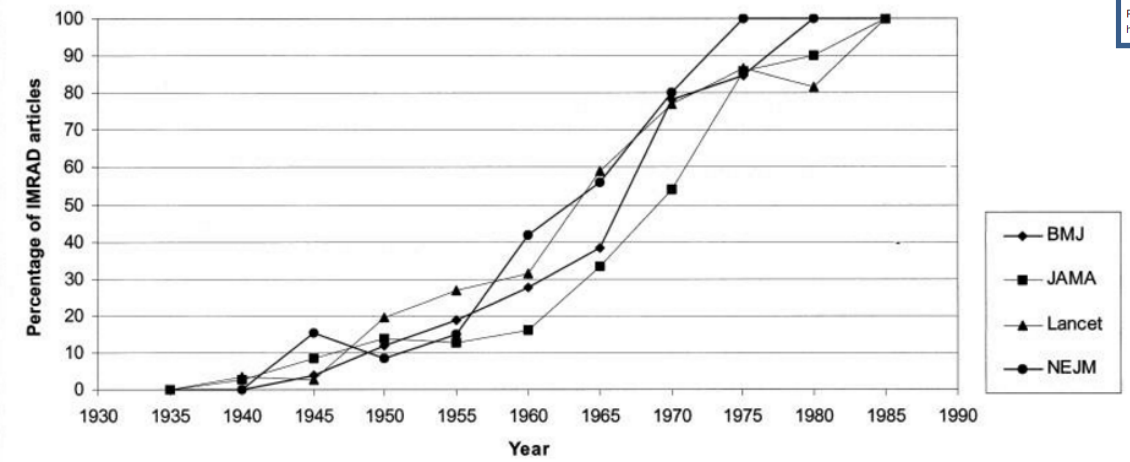
Annotation procedure and annotated corpus

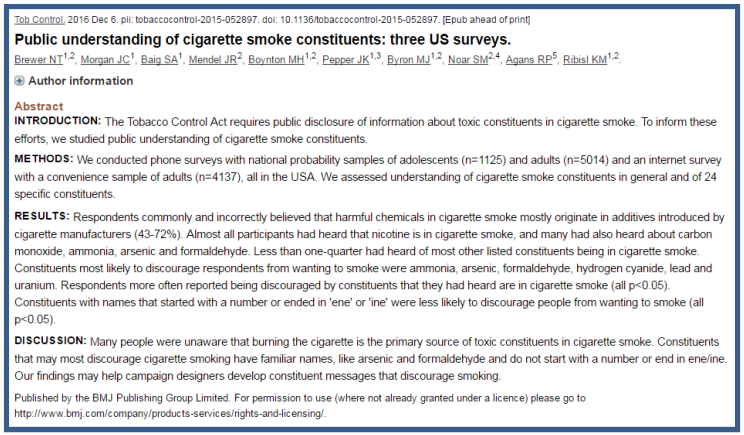
Zone Analysis: IMRAD

Luciana B. Sollaci & Mauricio G. Pereira (July 2004). *The introductio, methods, results, and discussion (IMRAD) structure: a fifty-year survey.* J Med Libr Assoc. 2004 July; 92(3): 364–371. 92 (3)

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

Random sample of (n = 1,297) articles published in *British Medical Journal, JAMA, The Lancet,* and the *New England Journal of Medicine,* 1935–1985





Tob Control. 2016 Dec 6. pii: tobaccocontrol-2015-052897. doi: 10.1136/tobaccocontrol-2015-052897. [Epub ahead of print]

Public understanding of cigarette smoke constituents: three US surveys.

Brewer NT[1,2], Morgan JC[1], Baig SA[1], Mendel JR[2], Boynton MH[1,2], Pepper JK[1,3], Byron MJ[1,2], Noar SM[2,4], Agans RP[5], Ribisl KM[1,2].

Author information

Abstract

INTRODUCTION: The Tobacco Control Act requires public disclosure of information about toxic constituents in cigarette smoke. To inform these efforts, we studied public understanding of cigarette smoke constituents.

METHODS: We conducted phone surveys with national probability samples of adolescents (n=1125) and adults (n=5014) and an internet survey with a convenience sample of adults (n=4137), all in the USA. We assessed understanding of cigarette smoke constituents in general and of 24 specific constituents.

RESULTS: Respondents commonly and incorrectly believed that harmful chemicals in cigarette smoke mostly originate in additives introduced by cigarette manufacturers (43-72%). Almost all participants had heard that nicotine is in cigarette smoke, and many had also heard about carbon monoxide, ammonia, arsenic and formaldehyde. Less than one-quarter had heard of most other listed constituents being in cigarette smoke. Constituents most likely to discourage respondents from wanting to smoke were ammonia, arsenic, formaldehyde, hydrogen cyanide, lead and uranium. Respondents more often reported being discouraged by constituents that they had heard are in cigarette smoke (all $p<0.05$). Constituents with names that started with a number or ended in 'ene' or 'ine' were less likely to discourage people from wanting to smoke (all $p<0.05$).

DISCUSSION: Many people were unaware that burning the cigarette is the primary source of toxic constituents in cigarette smoke. Constituents that may most discourage cigarette smoking have familiar names, like arsenic and formaldehyde and do not start with a number or end in ene/ine. Our findings may help campaign designers develop constituent messages that discourage smoking.

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

Scientific discourse characterization

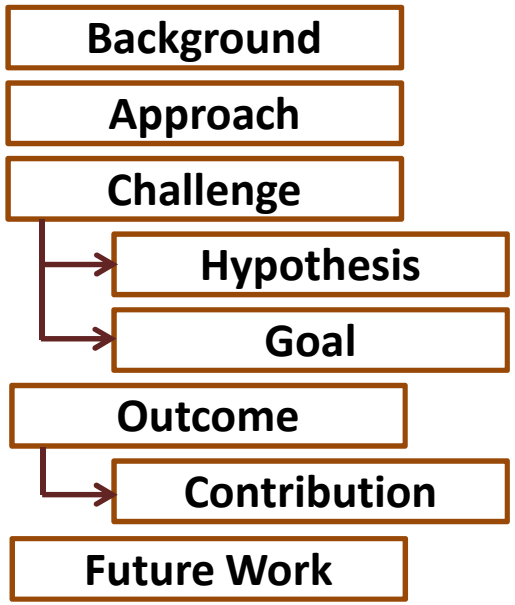
Annotation procedure and annotated corpus

Zone Analysis: Dr. Inventor Scientific Discurse Schema

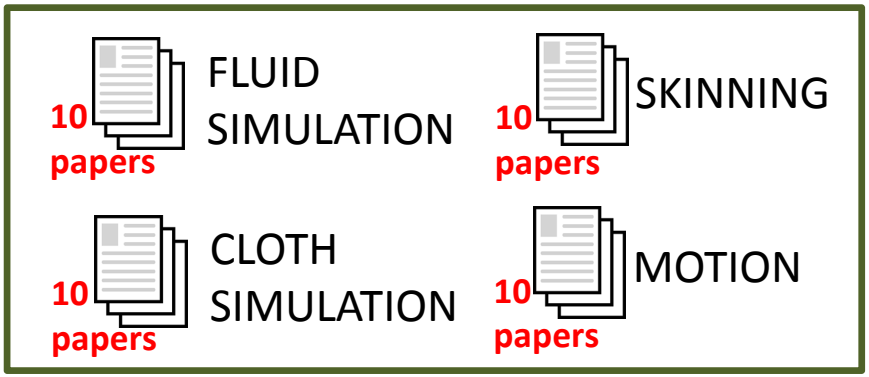
Fisas, B., Ronzano, F., & Saggion, H. (2015). *On the Discoursive Structure of Computer Graphics Research Papers.* In The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015 (p. 42).

Fisas, B., Ronzano, F., & Saggion, H. (2016). *A Multi-Layered Annotated Corpus of Scientific Papers.* LREC.

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



Dr. Inventor Corpus



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

Scientific discourse characterization

Annotation procedure and annotated corpus

Zone Analysis: Dr. Inventor Scientific Discurse Schema

Fisas, B., Ronzano, F., & Saggion, H. (2015). *On the Discoursive Structure of Computer Graphics Research Papers.* In The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015 (p. 42).

Fisas, B., Ronzano, F., & Saggion, H. (2016). *A Multi-Layered Annotated Corpus of Scientific Papers.* LREC.

Annotators

upf.

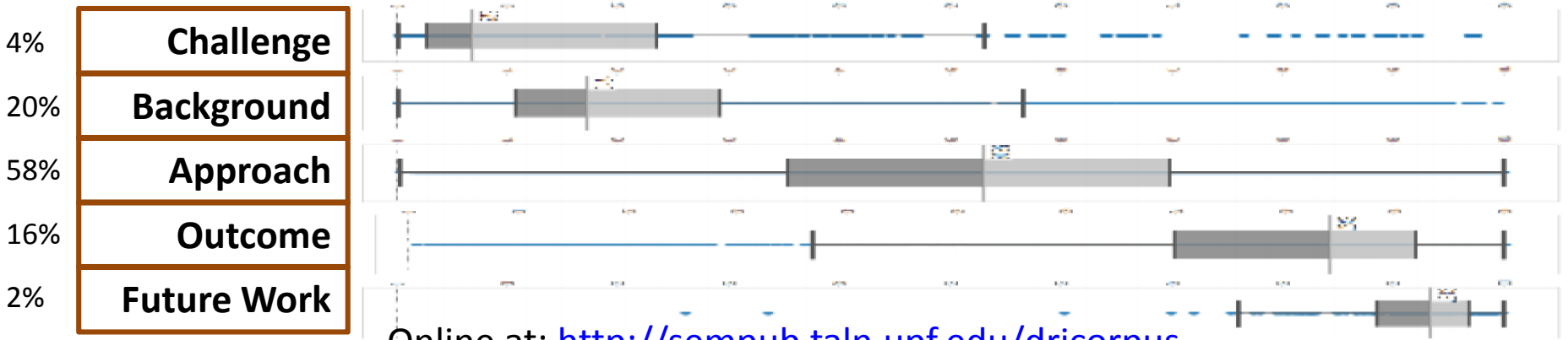
Avg. annotator agreement K: **0,67**

Annotation workflow

Training Session

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

Distribution of sentence rethorical class (over papers' length)



Online at: http://sempub.taln.upf.edu/dricorpus

Scientific discourse characterization

Automated annotation of scientific texts

Zone Analysis: Dr. Inventor Scientific Discurse Schema

Fisas, B., Ronzano, F., & Saggion, H. (2015). *On the Discoursive Structure of Computer Graphics Research Papers.* In The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015 (p. 42).

Fisas, B., Ronzano, F., & Saggion, H. (2016). *A Multi-Layered Annotated Corpus of Scientific Papers.* LREC.

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

	Category	**Logistic Regression**	**SVM**
58%	*Approach*	0.876	0.851
20%	*Background*	0.778	0.735
4%	*Challenge*	0.466	0.430
2%	*Future Work*	0.675	0.496
16%	*Outcome*	0.679	0.623
	Avg. F1:	**0.801**	**0.764**

Percentage of annotated sentences by category

Scientific discourse characterization

Automated annotation of ABSTRACTS of scientific texts

Guo, Y., Korhonen, A., Liakata, M., Karolinska, I. S., Sun, L., & Stenius, U. (2010, July). *Identifying the information structure of scientific abstracts: an investigation of three different schemes.* In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (pp. 99-107). ACL.

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

3 ANNOTATION SCHEMAS: (k measured over 1/3 of the corpus, three annotators)

- Objective, Method, Results and Conclusion (K=0,84) →SVM acc: 0.89
- AZ, 7 categories (K=0,85) →SVM acc: 0.90
- CoreSC 11 categories (K=0,50) →SVM 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)

Scientific discourse characterization

Automated annotation of ABSTRACTS of scientific texts

Hirohata, K., Okazaki, N., Ananiadou, S., Ishizuka, M., & Biocentre, M. I. (2008, January). *Identifying Sections in Scientific Abstracts using Conditional Random Fields.* In IJCNLP (pp. 381-388).

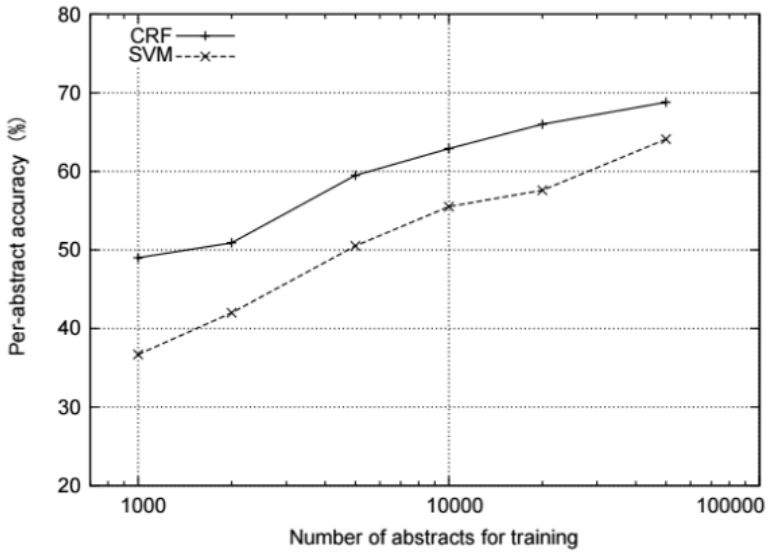
CORPUS: 51,000 MedLine abstracts with sentences divided in Objective, Method, Result and Conclusion

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



Scientific discourse characterization

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

Method	50	100	150	200	250	300	400	500
Random selection	0.73	0.76	0.77	0.78	0.78	0.78	0.79	0.79
Active learning								
Least confident	0.73	0.76	0.77	0.79	0.80	0.80	0.81	0.82
Margin	0.73	0.76	0.78	0.79	0.80	0.80	0.81	0.81
Query-by-bagging	0.73	0.77	0.78	0.78	0.79	0.79	0.80	0.81

Fully supervised accuracy (8,171 instances): 0.84

Scientific discourse characterization

Automated annotation of scientific texts

Guo, Y., Korhonen, A., & Poibeau, T. (2011, July). *A weakly-supervised approach to argumentative zoning of scientific documents.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 273-283). Association for Computational Linguistics.

- 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

Guo, Y., Reichart, R., & Korhonen, A. (2013, June). *Improved Information Structure Analysis of Scientific Documents Through Discourse and Lexical Constraints.* In HLT-NAACL (pp. 928-937).

- 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

Scientific discourse characterization

Automated annotation of scientific texts

Contractor, D., Guo, Y., & Korhonen, A. (2012). *Using Argumentative Zones for Extractive Summarization of Scientific Articles.* In COLING (Vol. 12, pp. 663-678).

- Annotation with AZ categories of a corpus of 50 biomedical articles sources from a number of journals on cancer
- Categories: Background, Conclusion, Problem, Connection, Method, Difference, Result and Future work
- Inter-annotator agreement $\kappa = 0.83$

Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010, July). *The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text.* In Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task (pp. 1-12). Association for Computational Linguistics.

Lin, J., Karakos, D., Demner-Fushman, D., & Khudanpur, S. (2006, June). *Generative content models for structural analysis of medical abstracts.* In Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology (pp. 65-72). Association for Computational Linguistics.

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
https://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/

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 onCancer Risk Assessment
http://www.cl.cam.ac.uk/~yg244/abstract_az.html

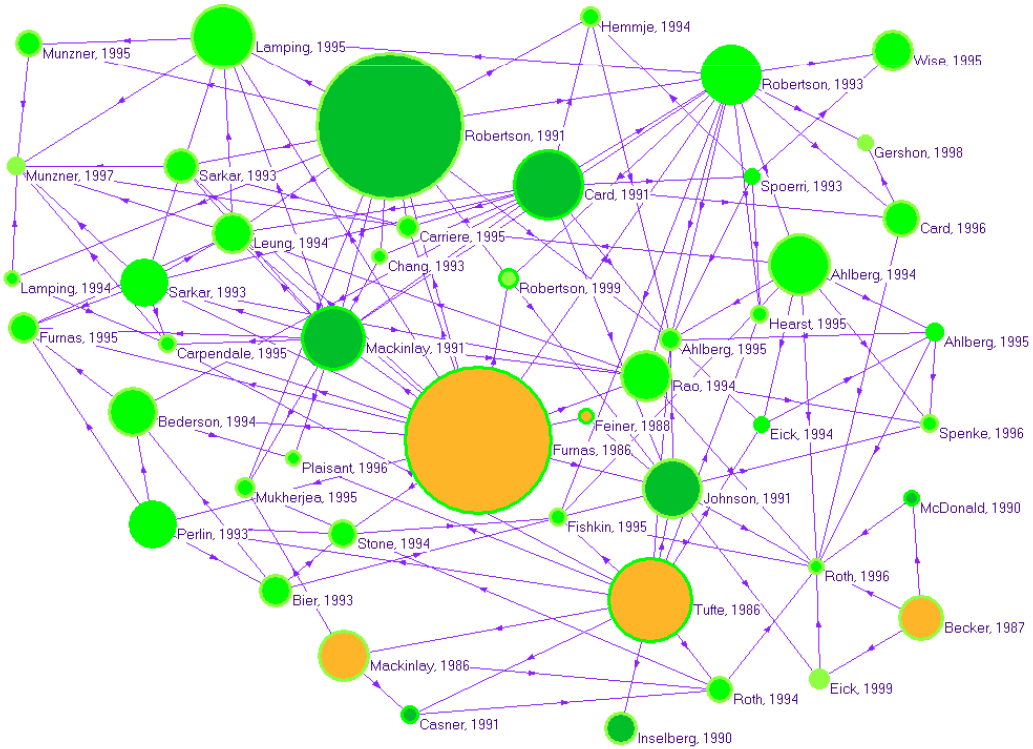
Conclusions

- 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

Universitat Pompeu Fabra Barcelona

EXCELENCIA MARÍA DE MAEZTU

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.

random variables. Sung [8] improved the result of Cai [6] for NA random variables under much weaker conditions.

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

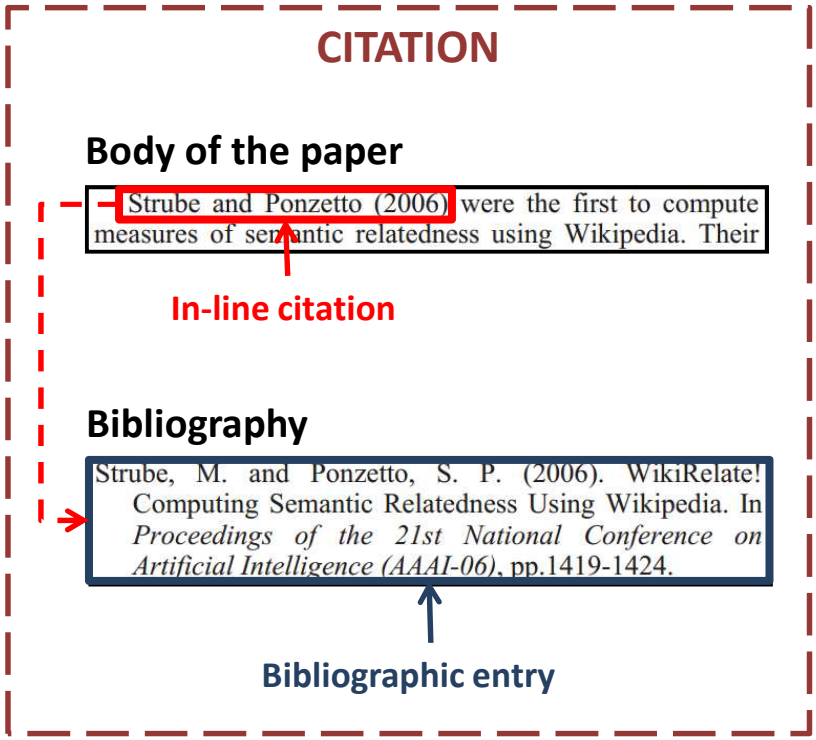
- Citations in scientific literature
- How citations are studied?
  - Citation network analysis
  - Citation function
  - Citation prediction and recommendation
  - Citation-based summarization
- Citation graphs
- Conclusions

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:

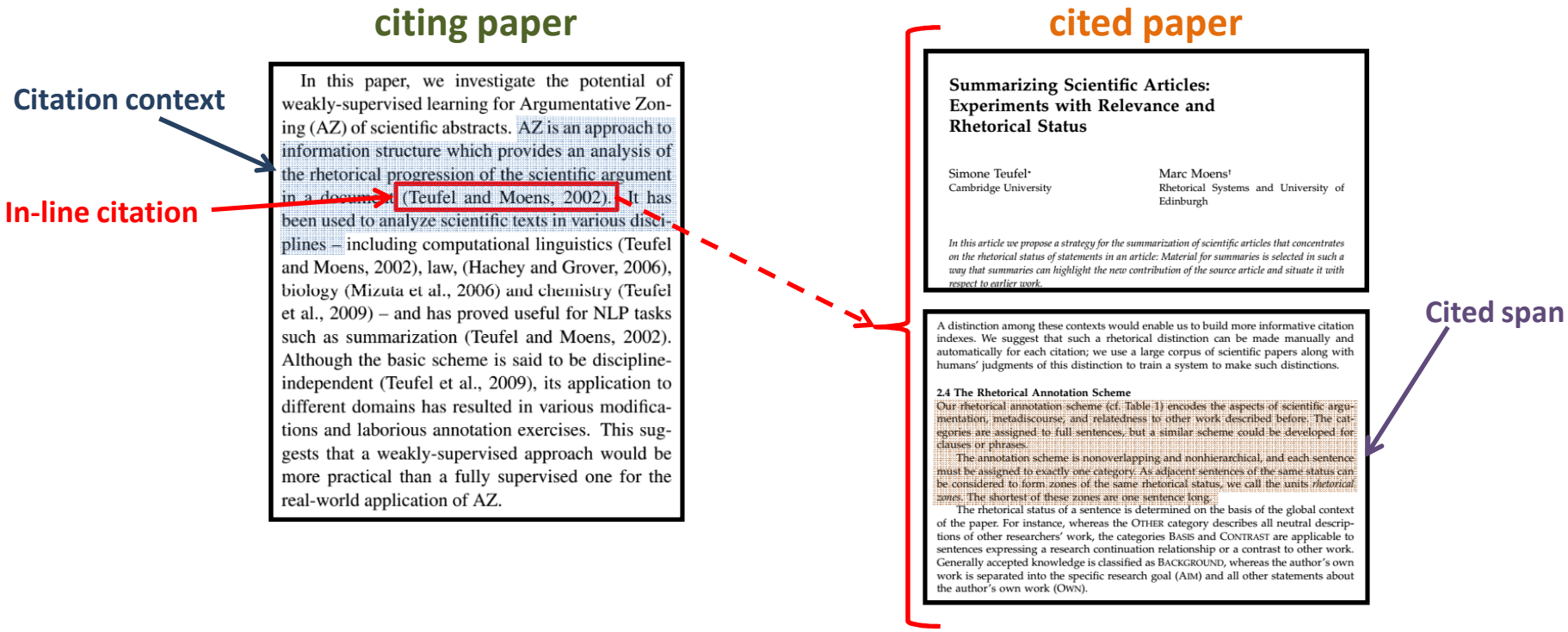
- **ground the arguments** and give the work **factual basis**
- **avoid plagiarism** (intellectual honesty)
- **attribute prior or unoriginal work and ideas** to the correct sources
- allow the reader to **determine independently whether the referenced material supports the author's argument** in the 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



Citations in scientific literature

The elements of a citation

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



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

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

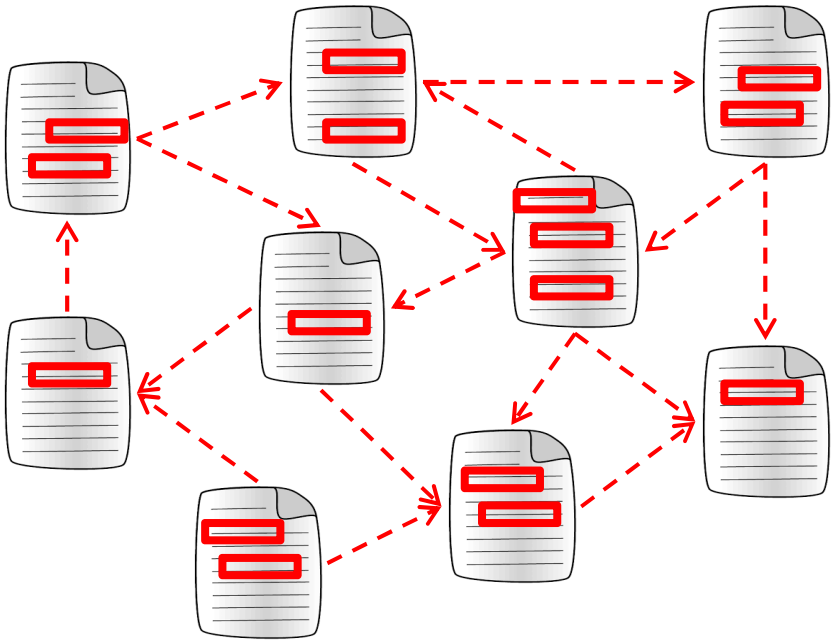
Citation network analysis

Citation networks

Citation networks:

- ***nodes*: papers**
- ***arks*: directed from citing to cited paper**

The more often a single paper is cited, the more important it seems to be



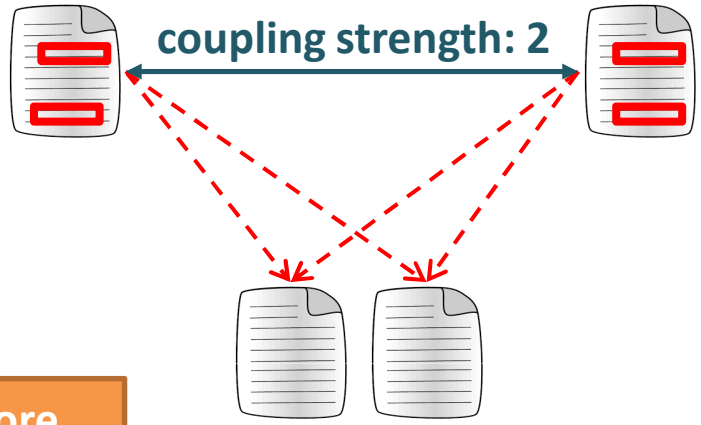
Citation network analysis

Bibliographic coupling and co-citation networks

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



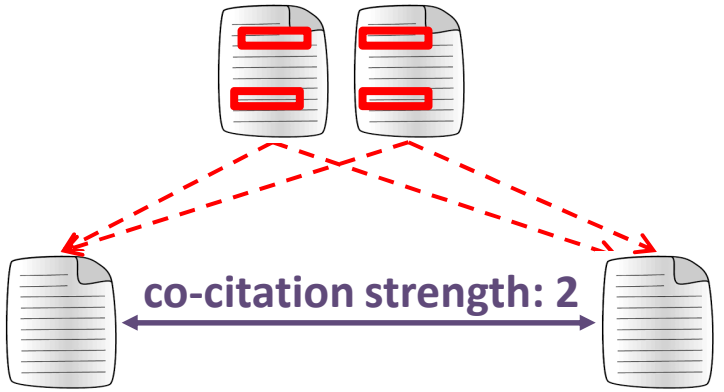
The more often two papers are cited together, the more likely they are to be part of some research question or ongoing problem or conversation topic within the discipline

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.



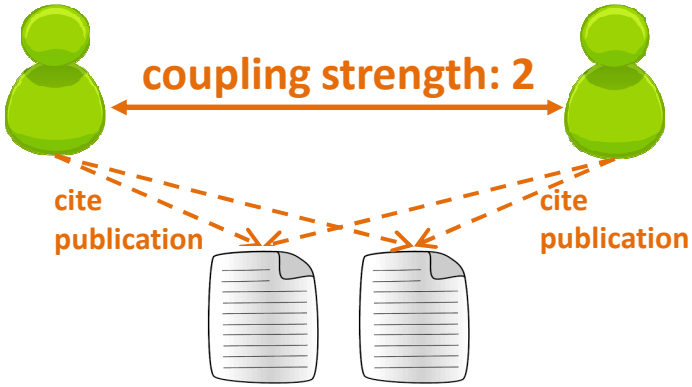
INTERACTIVE EXAMPLE:
http://jgoodwin.net/network/cites-slider.html

Citation network analysis

Author coupling and co-citation networks

Author bibliographic coupling:

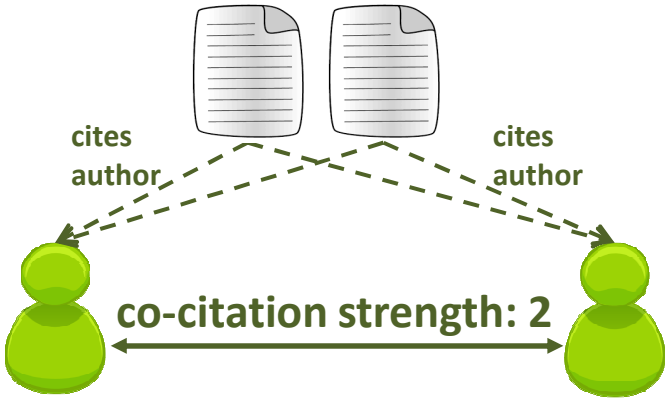
- ***nodes***: authors
- ***arks***: undirected, equals to the number of references that the publications of the pairs of authors have in common



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 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

Citation networks

Microsoft Academic Graph

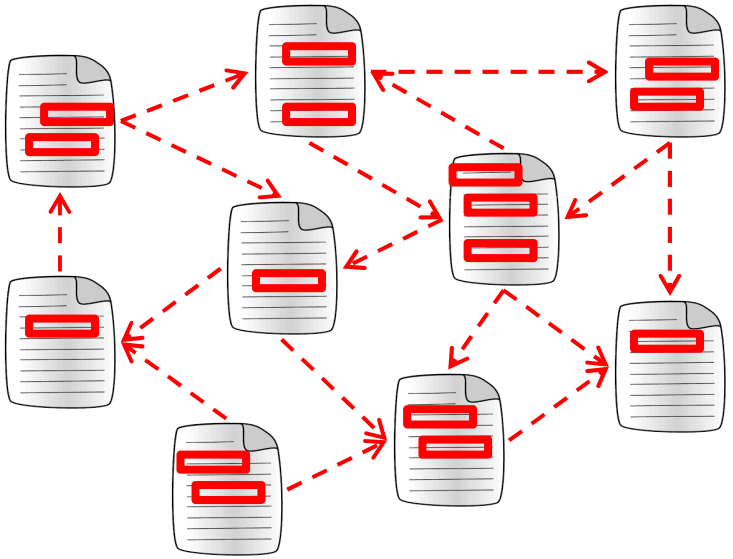
February 2016

- 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^{x} 10M

October 2016

- 384,413 papers
- 1,751,463 internal citations
- each paper in the graph is cited on average 4.56 times

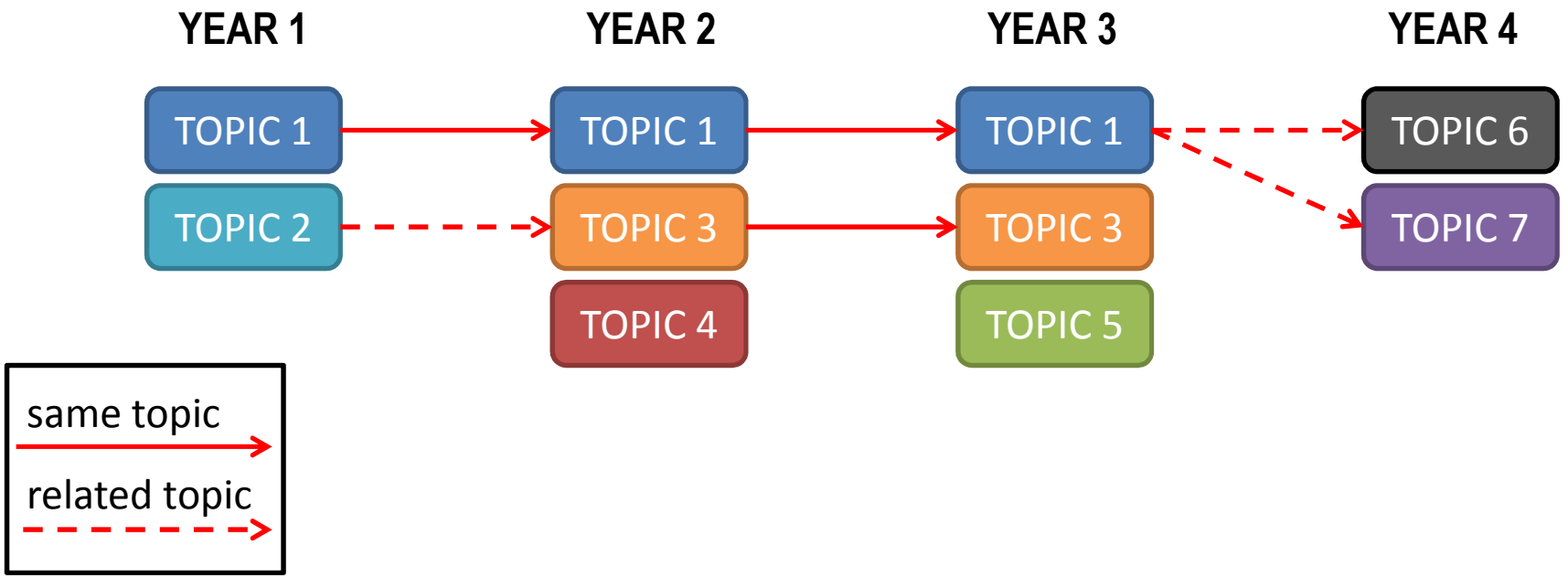


Citation network analysis

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



Citation network analysis

Improving detection of scientific topic evolution by citation network

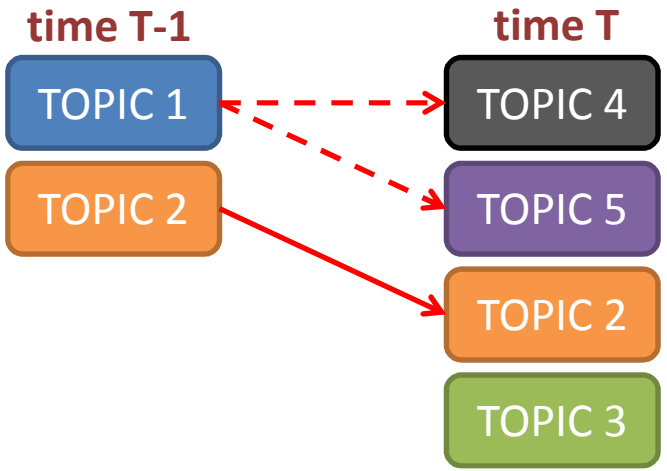
Once defined a number **k** of topics, given a collection of documents, a topic detection method generates for each topic **z** a vocabulary distribution so as to maximize the likelihood of the observed data

Topic	Vocabulary distribution		
TOPIC 1	[Word 1: ***0,4***	Word 2: ***0,2***	Word 3: ***0,4***]
TOPIC 2	[Word 1: ***0,7***	Word 2: ***0,1***	Word 3: ***0,2***]

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 sim(**z(T)**, **z(T-1)**)



He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009, November). *Detecting topic evolution in scientific literature: how can citations help?*. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 957-966). ACM.

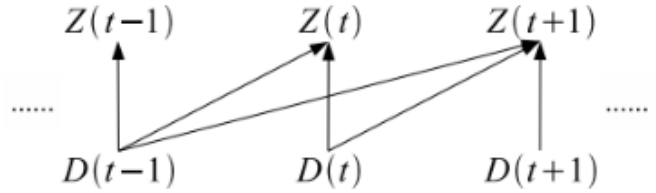
Citation network analysis

Improving detection of scientific topic evolution by citation network

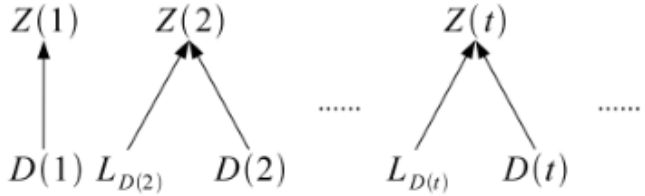
Considering a collection of scientific paper spanning a number of years, in order to track year-by-year topic evolution, we can generate the topic of each year by different approaches:

- **Time independent topic evolution learning**

- **Accumulative topic evolution learning**



- **Citation-aware topic evolution learning**

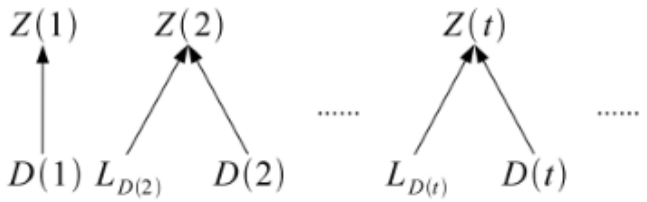


He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009, November). *Detecting topic evolution in scientific literature: how can citations help?.* In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 957-966). ACM.

Citation network analysis

Improving detection of scientific topic evolution by citation network

Citation-aware topic evolution learning



DRAWBACKS

- not all citations are equally important (only few can ne 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

$$Z(t) = \arg\max_{Z(t)} \prod_{d \in D(t) \cup L_{D(t)}} p'(d|Z(t)),$$

where

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

Citing paper — Autonomous part (new ideas)

Cited papers — Inherited part (previous work)

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

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009, November). *Detecting topic evolution in scientific literature: how can citations help?.* In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 957-966). ACM.

Citation network analysis

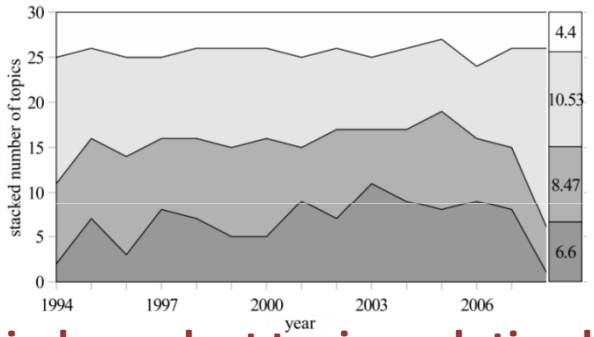
Improving detection of scientific topic evolution by citation network

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

Evolution of 30 topics studied with different approaches

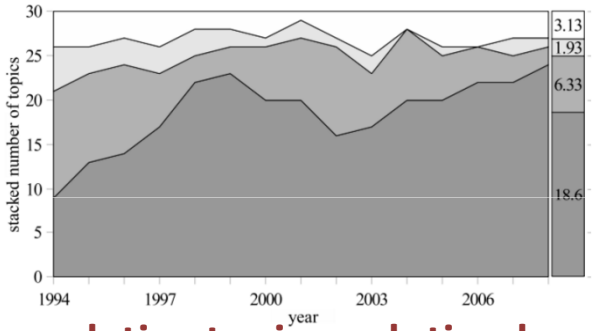
noisy topics new topics similar topics same topics

Citation unaware



Large number of new topics. Several noisy topics

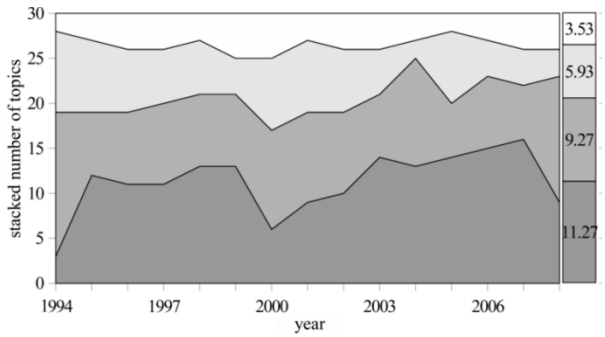
Time independent topic evolution learning



Historical topics tends to dominate; difficult to detect new ones

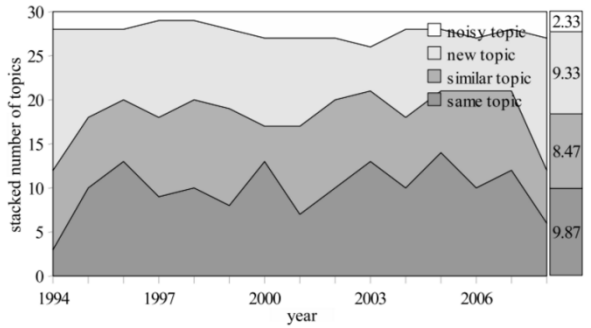
Accumulative topic evolution learning

Citation aware



Produce less new topics with respect a ITM since old, cited papers are treated as new ones

Citation-aware topic evolution learning



Reach a good balance between new and old topics enabling detection of new ones

Inheritance topic model

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009, November). *Detecting topic evolution in scientific literature: how can citations help?.* In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 957-966). ACM.

Citation network analysis

Co-authorship network

reflects the personal link between scientists

Dataset: 2 co-authorship networks (1991-1998):

- **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**

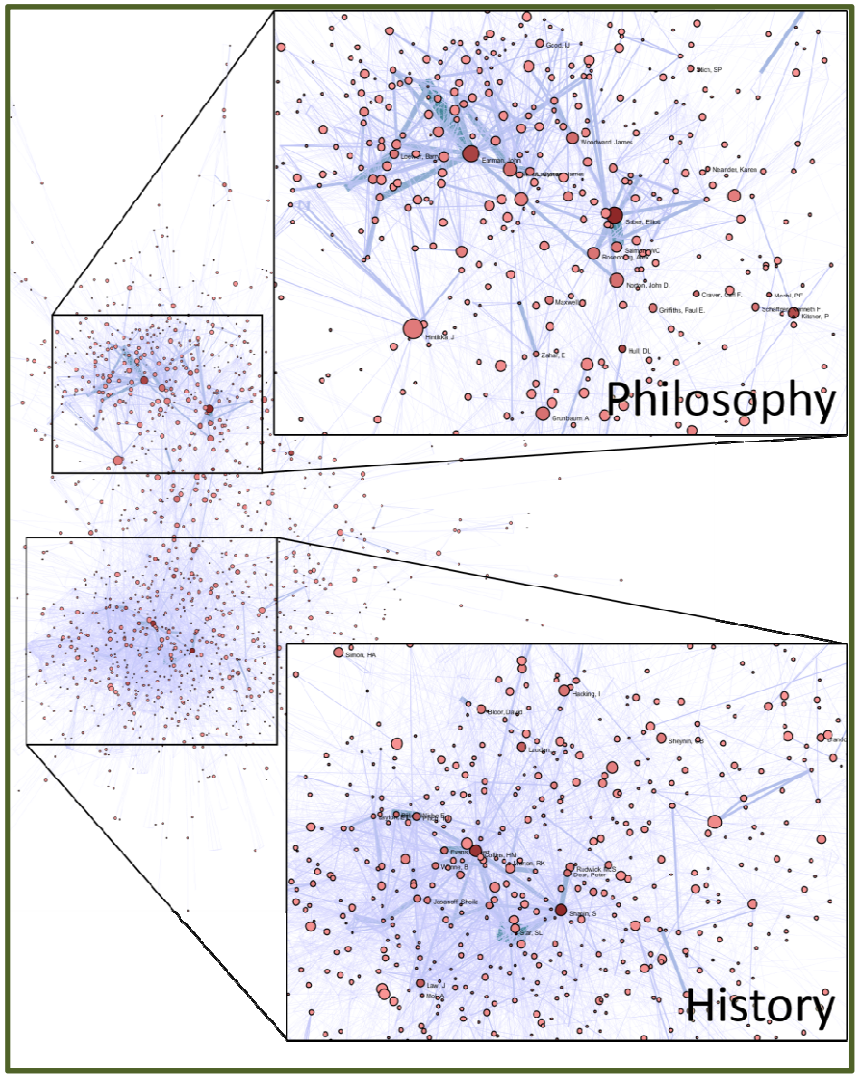
Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). *Evolution of the social network of scientific collaborations.* Physica A: Statistical mechanics and its applications, 311(3), 590-614.

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 function

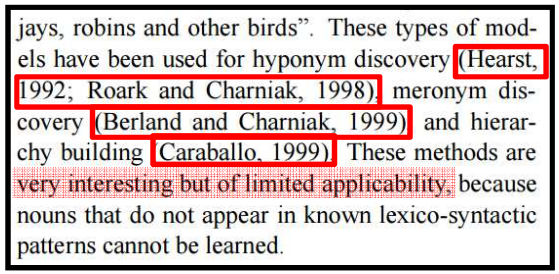
Many research impact and quality indexes are based on citation counts but…

Not all citations are equal!

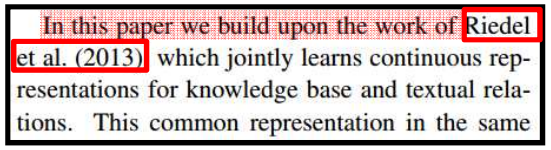
$$\text{Impact Factor (corrected)} = \frac{\text{\# times your work is cited} - \text{\# citations that actually trash your work} - \text{\# times you cited yourself (nice try)} - \text{\# times you were cited just to pad the introduction section} - \text{\# citations the editor pressured the author to include to increase the journal's impact factor}}{\text{\# original articles you've written} + \text{\# articles you were included in out of pity or politics} + \text{\# not-so-original articles you've ~~written~~ copied and pasted}}$$

JORGE CHAM © 2008

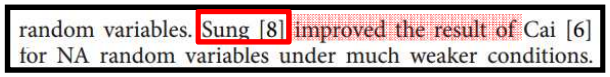
There are different motivation that could explain why an author cites other pieces of research


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.

criticize a work express **contrary or negative judgments**


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

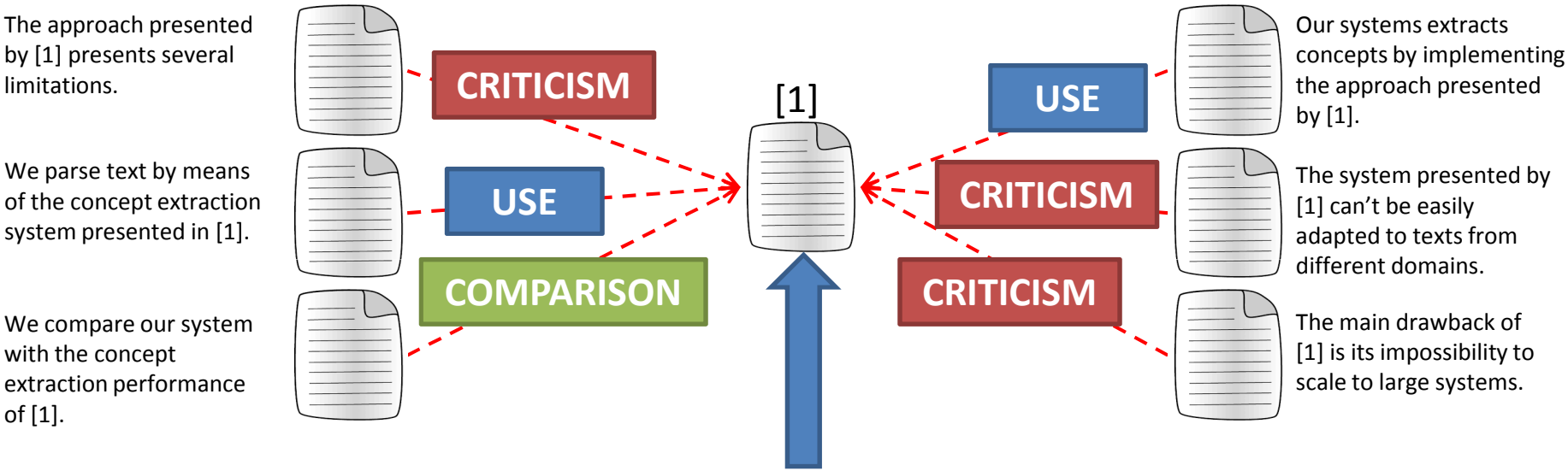

random variables. Sung [8] improved the result of Cai [6] for NA random variables under much weaker conditions.

highlight a **positive result**

Citation function

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

Not all citations are equal!



This paper has 6 citations!

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

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

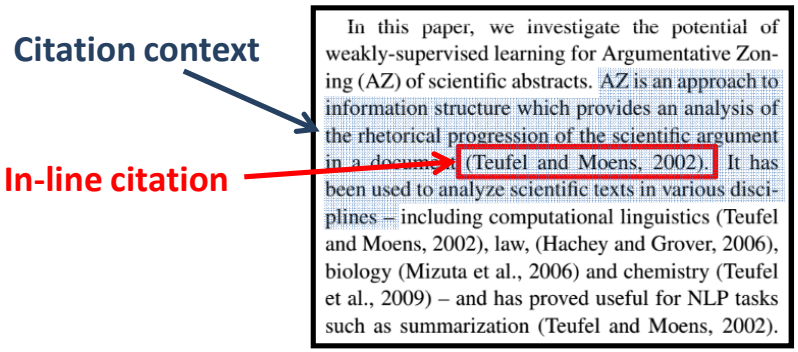
COMPARISON One citation of this paper **compares** the approach / tool presented.

Citation function

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

Citation function

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

Moravcsik, M. J., & Murugesan, P. (1975). *Some results on the function and quality of citations.* Social studies of science, 5(1), 86-92.

4 dimensions for citation characterization

Conceptual	*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.*	**use of theory**
Operational	*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.*	**use of technical method**
Organic	*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.*	**work is crucially needed for understanding of citing article**
Perfunctionary	*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.*	**just a general acknowledgement**
Evolutionary	*[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.*	**own work is an alternative to cited work**
Juxtapositional	*[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.*	**own work is based on the cited work**
Confirmative	*A reference is confirmative if the author of the citing paper considers the paper referred to as correct.*	**the work confirm the cited paper**
Negative	*The author of the citing paper is not certain about the correctness of the cited paper.*	**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

Spiegel-Rösing, I. (1977). *Science studies: Bibliometric and content analysis.* Social Studies of Science, 97-113.

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

(2,309 citations from Science Studies Vol. 1-4 1971-1974)

Citation function

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

Teufel, S., Siddharthan, A., & Tidhar, D. (2009, July). *An annotation scheme for citation function.* In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (pp. 80-87). ACL

4 top level categories for citation characterization

Weakness		*Authors point out a weakness in cited work.*
Contrast		*Authors make contrast/comparison with cited work (4 categories)*
	CoCoGM	*Contrast/Comparison in Goals or Methods (neutral)*
	CoCoR0	*Contrast/Comparison in Results (neutral)*
	CoCo	*Unfavourable Contrast/Comparison (current work is better than cited work)*
	CoCoXY	*Contrast between 2 cited methods*
Positive		*Authors agree with/make use of/show compatibility or similarity with cited work (6 categories),*
	PBas	*author uses cited work as starting point*
	PUse	*author uses tools/algorithms/data*
	PModi	*author adapts or modi_es tools/algorithms/data*
	PMot	*this citation is positive about approach or problem addressed (used to motivate work in current paper)*
	PSim	*author's work and cited work are similar*
	PSup	*author's work and cited work are compatible/ provide support for each other*
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

Neut	PUse	CoCoGM	PSim	Weak	CoCoXY	PMot	PModi	PBas	PSup	CoCo-	CoCoR0
62.7%	15.8%	3.9%	3.8%	3.1%	2.9%	2.2%	1.6%	1.5%	1.1%	1.0%	0.8%

Online at: http://www.cl.cam.ac.uk/~sht25/Project_Index/Citraz_Index.html http://www.cl.cam.ac.uk/~sht25/CFC.html

Citation function

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

Teufel, S., Siddharthan, A., & Tidhar, D. (2009, July). *An annotation scheme for citation function.* In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (pp. 80-87). ACL

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

	Weakness	Positive	Contrast	Neutral
P	.80	.75	.77	.81
R	.49	.65	.52	.90
F	.61	.70	.62	.86

All classes

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

Citation function

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

Abu-Jbara, A., Ezra, J., & Radev, D. R. (2013). *Purpose and Polarity of Citation: Towards NLP-based Bibliometrics.* In HLT-NAACL (pp. 596-606).

6 classes for citation characterization

Criticism	*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.*
Comparison	*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.*
Use	*A citing sentence is classified as "Use" when the citing paper uses the method, idea or tool of the cited paper.*
Substantiation	*A citing sentence is classified as "Substantiating" when the results, claims of the citing work substantiate, verify the cited paper and support each other.*
Basis	*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.*
Neutral	*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.*

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

Citation function

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

Abu-Jbara, A., Ezra, J., & Radev, D. R. (2013). *Purpose and Polarity of Citation: Towards NLP-based Bibliometrics.* In HLT-NAACL (pp. 596-606).

Classifiers: CRF

Citation context features: (ordered by relevance)

Feature	Description
Demonstrative determiners	Takes a value of 1 if the current sentence contains contains a *demonstrative determiner* (this, these, etc.), and 0 otherwise.
Conjunctive adverbs	Takes a value of 1 if the current sentence starts with a *conjunctive adverb* (However, Furthermore, Accordingly, etc.), and 0 otherwise.
Position	Position of the current sentence with respect to the citing sentence. This feature takes one of four values: -1, 0, 1, and 2.
Contains Closest Noun Phrase	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.
2-3 grams Contains Other references	The first bigram and trigram in the sentence (*This approach*, *One problem with*, etc.). Takes a value of 1 if the current sentence contains references other than the target, and 0 otherwise.
Contains a Mention of target reference	Takes a value of 1 if the current sentence contains a mention (explicit or anaphoric) of the target reference, and 0 otherwise.
Multiple references	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.

Results:

	Precision	Recall	F1
CRFs	**98.5%**	**82.0%**	**89.5%**
ALL	30.7%	100.0%	46.9%
CS-ONLY	88.0%	74.0%	80.4%
SVM	92.0%	76.4%	83.5%

Lexical features are more important than structural features

Citation function

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

Abu-Jbara, A., Ezra, J., & Radev, D. R. (2013). *Purpose and Polarity of Citation: Towards NLP-based Bibliometrics.* In HLT-NAACL (pp. 596-606).

Classifiers: SVM with linear kernel

Citation features: (ordered by relevance)

Feature	Description
Reference count	The number of references that appear in the citation context.
Is Separate	Whether the target reference appears within a group of references or separate (i.e. single reference).
Closest Verb / Adjective / Adverb	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.
Self Citation	Whether the citation from the source paper to the target reference is a self citation.
Contains 1st/3rd PP	Whether the citation context contains a first/third person pronoun.
Negation	Whether the citation context contains a negation cue. The list of negation cues is taken from the training data of the *SEM 2012 negation detection shared task (Morante and Blanco, 2012).
Speculation	Whether the citation context contains a speculation cue. The list is taken from Quirk et al. (1985)
Closest Subjectivity Cue	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)
Contrary Expressions	Whether the citation context contains a contrary expression. The list is taken from Biber (1988)
Section	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.
Dependency Relations	All the dependency relations that appear in the citation context. For example, $nsubj(outperform, algorithm)$ is one of the relations extracted from "This algorithm outperforms the one proposed by...". 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).

Results:

	Criticism	Comparison	Use	Substantiating	Basis	Other
Precision	53.0%	55.2%	60.0%	50.1%	47.3%	64.0%
Recall	77.4%	43.1%	73.0%	57.3%	39.1%	85.1%
F1	63.0%	48.4%	66.0%	53.5%	42.1%	73.1%
Accuracy: 70.5%						
Macro-F: 58.0%						

- 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)

Citation function

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

Athar, A. (2011, June). *Sentiment analysis of citations using sentence structure-based features.* In Proceedings of the ACL 2011 student session (pp. 81-87). Association for Computational Linguistics.

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 function

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

Athar, A. (2011, June). *Sentiment analysis of citations using sentence structure-based features.* In Proceedings of the ACL 2011 student session (pp. 81-87). Association for Computational Linguistics.

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/

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

Athar, A., & Teufel, S. (2012, July). Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse* (pp. 18-26). Association for Computational Linguistics. → **SVM based approach to identify sentences belonging to the citation context. The information from the citation context improve sentiment analysis performance for citations**

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

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

Citation function

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

Shotton, D. (2010). *CiTO, the citation typing ontology.* Journal of biomedical semantics, 1(1), 1.

23 properties for citation characterization

Factual relationships	Rhetorical relationships		
	Positive	Negative	Neutral
cito:cites	*cito:confirms*	*cito:corrects*	*cito:discusses*
cito:citesAsAuthority	*cito:credits*	*cito:critiques*	*cito:reviews*
cito:citesAsMetadataDocument	*cito:extends*	*cito:disagreesWith*	
cito:citesAsSourceDocument	*cito:obtainsSupportFrom*	*cito:qualifies*	
cito:citesForInformation	*cito:supports*	*cito:refutes*	
cito:isCitedBy	*cito:updates*		
cito:obtainsBackgroundFrom			
cito:sharesAuthorsWith			
cito:usesDataFrom			
cito:usesMethodIn			

Part of the **Semantic Publishing and Referencing Ontologies**, includes 41 properties for citation characterization in its most recent version (03/07/2015)

Citation function

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

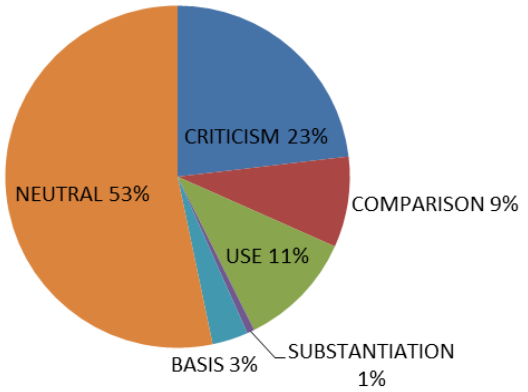
Fisas, B., Ronzano, F., & Saggion, H. (2016). *A Multi-Layered Annotated Corpus of Scientific Papers.* LREC.

6 top-level purposes and 16 sub-purposes for citation characterization

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.
	OTHER	
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 Grahics articles
- 1,575 citations



Online at:
http://sempub.taln.upf.edu/dricorpus

Citation function

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

Valenzuela, M., Ha, V., & Etzioni, O. (2015, April). *Identifying meaningful citations.* In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

Coarse and fine-grained labels for citation type

Citation Type	Fine-grained Label	Coarse Label
Related work	0	Incidental
Comparison	1	Incidental
Using the work	2	Important
Extending the work	3	Important

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

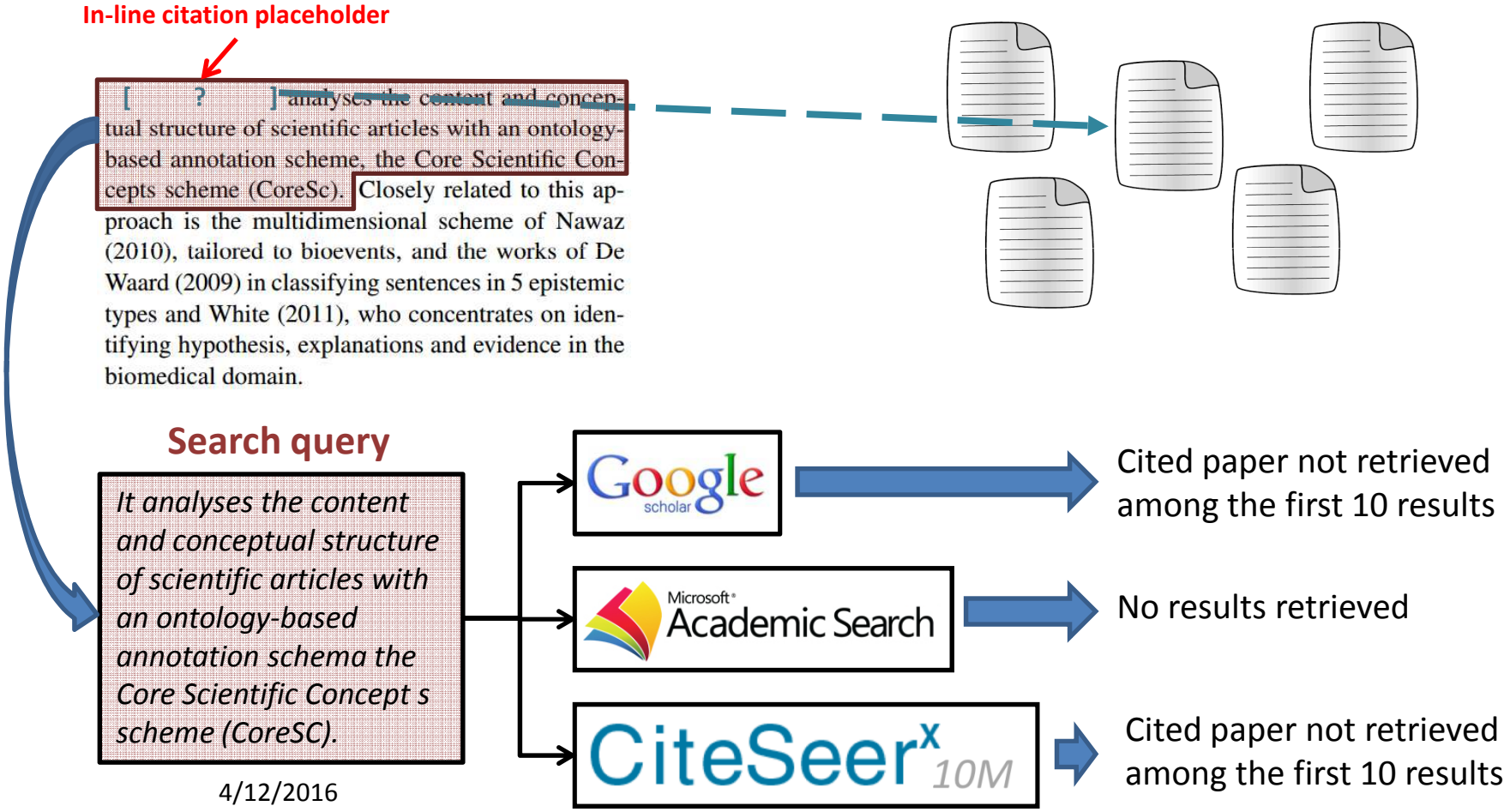
Algorithms: SVM (RBF k.) and random forest

Evaluation: SVM accuracy: 0,93

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**

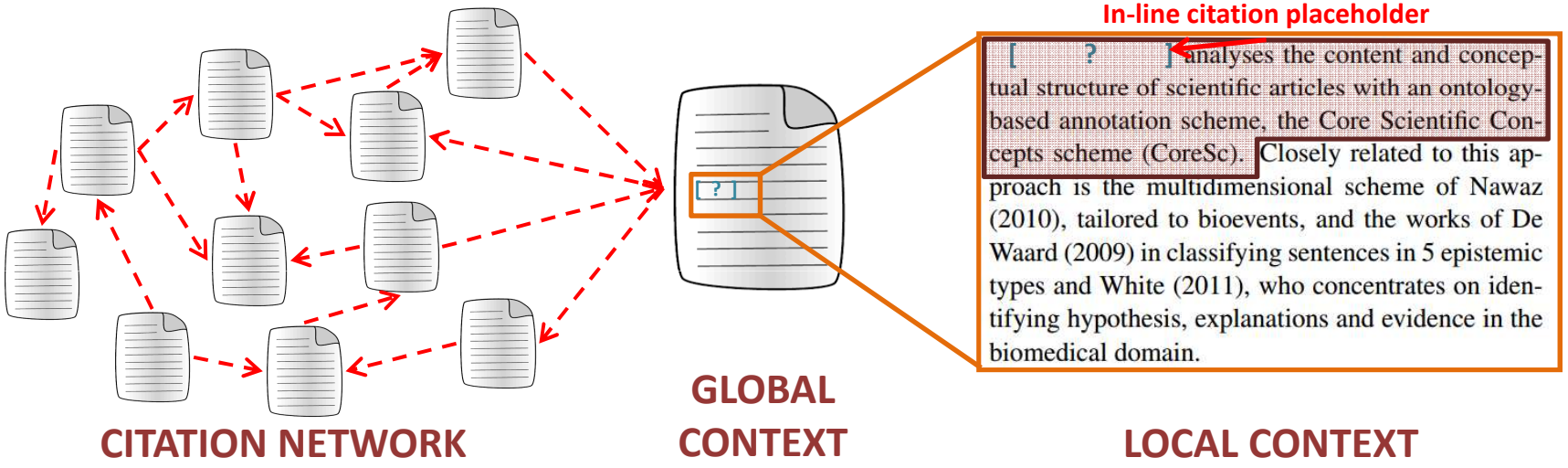
Citation prediction and recommendation

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



Citation prediction and recommendation

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



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…

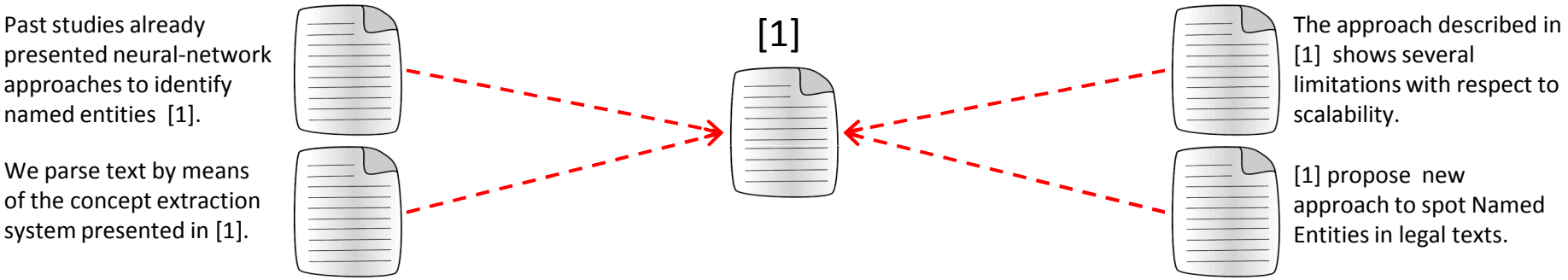
- **Huge search space**: progressive reduction of candidate set (lightweight cited papers selection methods for a first coarse-grained selection, candidate cited paper clustering)
- **Neural models**: estimate the probability that, given a word from the citation context, a document is cited by jointly 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

RefSeer

Online tool that implements distinct approaches of document level and context level citation recommendation - http://refseer.ist.psu.edu/

Citation-based summarization

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.

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 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)
https://psu.app.box.com/v/refseer (2015 dataset)

ACL Anthology Network

21,212 papers with 110,976 citations (Dec. 2013)
http://clair.eecs.umich.edu/aan/index.php (2013 dataset)

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

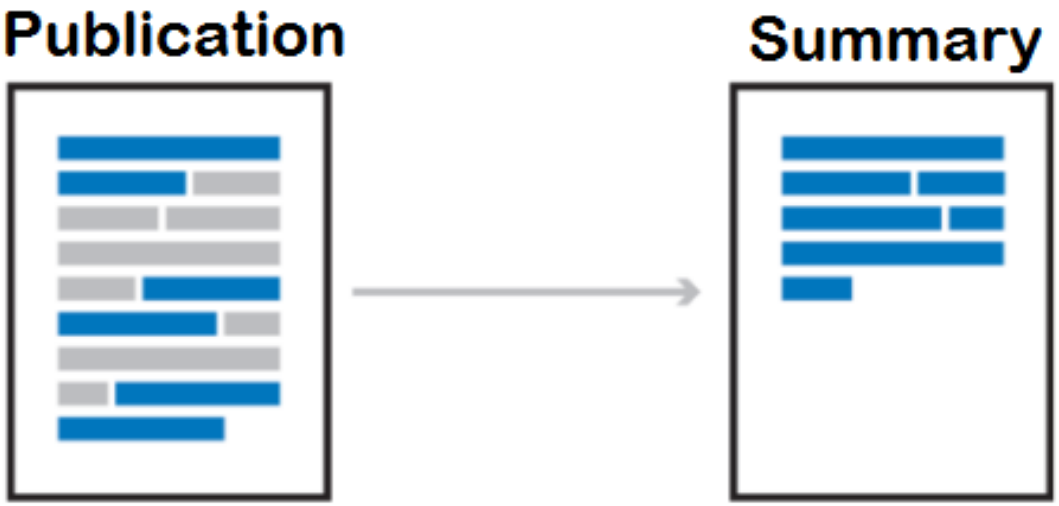
- Citations represent a primary device of scientific literature useful to issue **explicit author-created links among publications**
- Both the network of citations and the textual contents of citation contexts are exploited in many different tasks including: **research collaboration analysis**, **topic 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

Universitat Pompeu Fabra Barcelona

taln

EXCELENCIA MARÍA DE MAEZTU

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

Document summarization overview

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**.

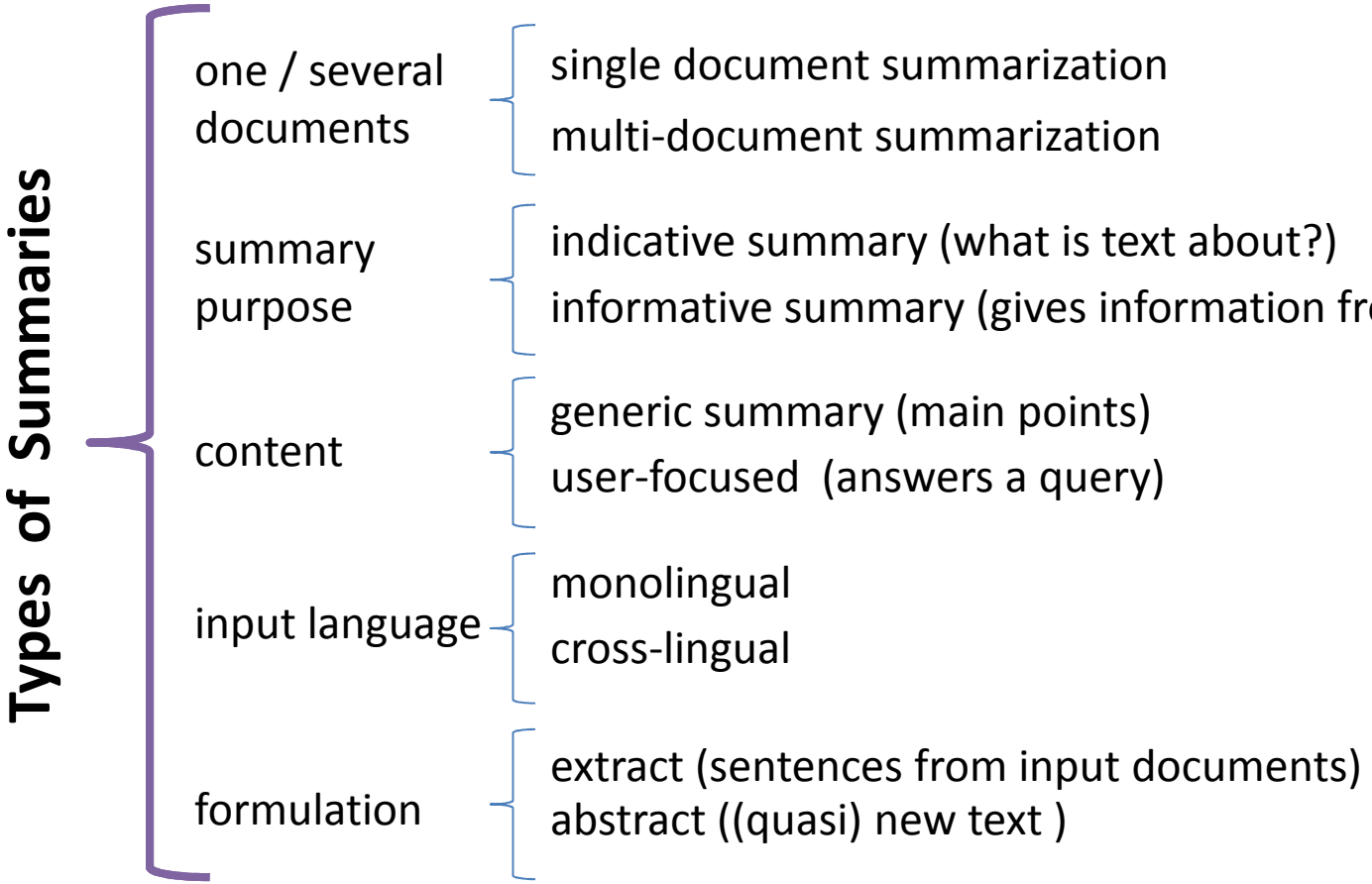
What is a summarizer?

An **algorithm** that **selects and presents**
the most important content of a document

Document summarization overview

Different types of summaries

Summaries should take into account a number of input factors such as the audience/reader of the summary



Spark Jones (2007)

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

Document summarization overview

Summarization by sentence extraction: sentence relevance

Function to **assess the contribution of a sentence to a summary**, developed since the late 50s some still used in the literature

- **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

Keyword method

Hypothesis: word/term repetition in a document taken as a measure of relevance of a word, however it is not enough

- Text Processing
  - stemming can be used for word normalization
  - stop word list can be used to filter non-content words
- Inverse document frequencies (over large corpora) should be used to assess word relevance

$$relevance(t) = tf(t) * idf(t)$$

$$idf(term) = \log\left(\frac{NUMDOC}{NUMDOC(term)}\right)$$

(Luhn, 1958)

Keyword method

Sentence score is proportional to the scores of clusters of keywords it contains

- Keywords are words w such that tf(w)*idf(w) > threshold
- "clusters" of *keywords* are identified in each sentence (e.g $[X_i \ldots X_{i+1} \ldots X_{i+n-1}]$ where X_i are keywords)
- Cluster weights: $w(C) = \frac{\#significant(C)^2}{\#words(C)}$
- Sentence score: $w(s) = \sum_{c \in s} w(c)$
- Term frequency is still used in many text summarization approaches, although its effectiveness is sometimes questioned

(Luhn, 1958)

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

$$relevance(t) = tf(t) * idf(t) \qquad idf(term) = \log\left(\frac{NUMDOC}{NUMDOC(term)}\right)$$

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)

Document summarization overview

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)

Document summarization overview

Summarization by sentence extraction: feature combinations

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

$$Weight(S) = \alpha.Title(S) + \beta.Cue(S) + \gamma.Keyword(S) + \delta.Position(S)$$

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

(Edmundson, 1969)(Kupiec et al, 1995)

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

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 (s1, s2, ..., sn)
- There are edges E(*sk*, *sl*) connecting *sk* with *sl*
- In(*si*) is the set of of sentences *si* such that there is an edge E(*si*, *sj*)
- Out(*si*) is the set of of sentences *si* such that there is an edge E(*sj*, *si*)
- Graph **generally undirected but could be directed if text order is taken into account** (*si* connects with *sj* only if $i < j$)

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

$$w(si) = \frac{d}{N} + (1-d) * \sum_{sj \in In(si)} \frac{sim(si,sj)}{\sum_{sk \in Out(sj)} sim(sk,sj)} * w(sj)$$

TextRank

$$w(si) = (1-d) + d * \sum_{sj \in In(si)} \frac{wj,i}{\sum_{sk \in Out(sj)} wj,k} * w(sj)$$

$$wj,k = \frac{|sj \cap sk|}{\log(|sj|)+\log(|sk|)}$$

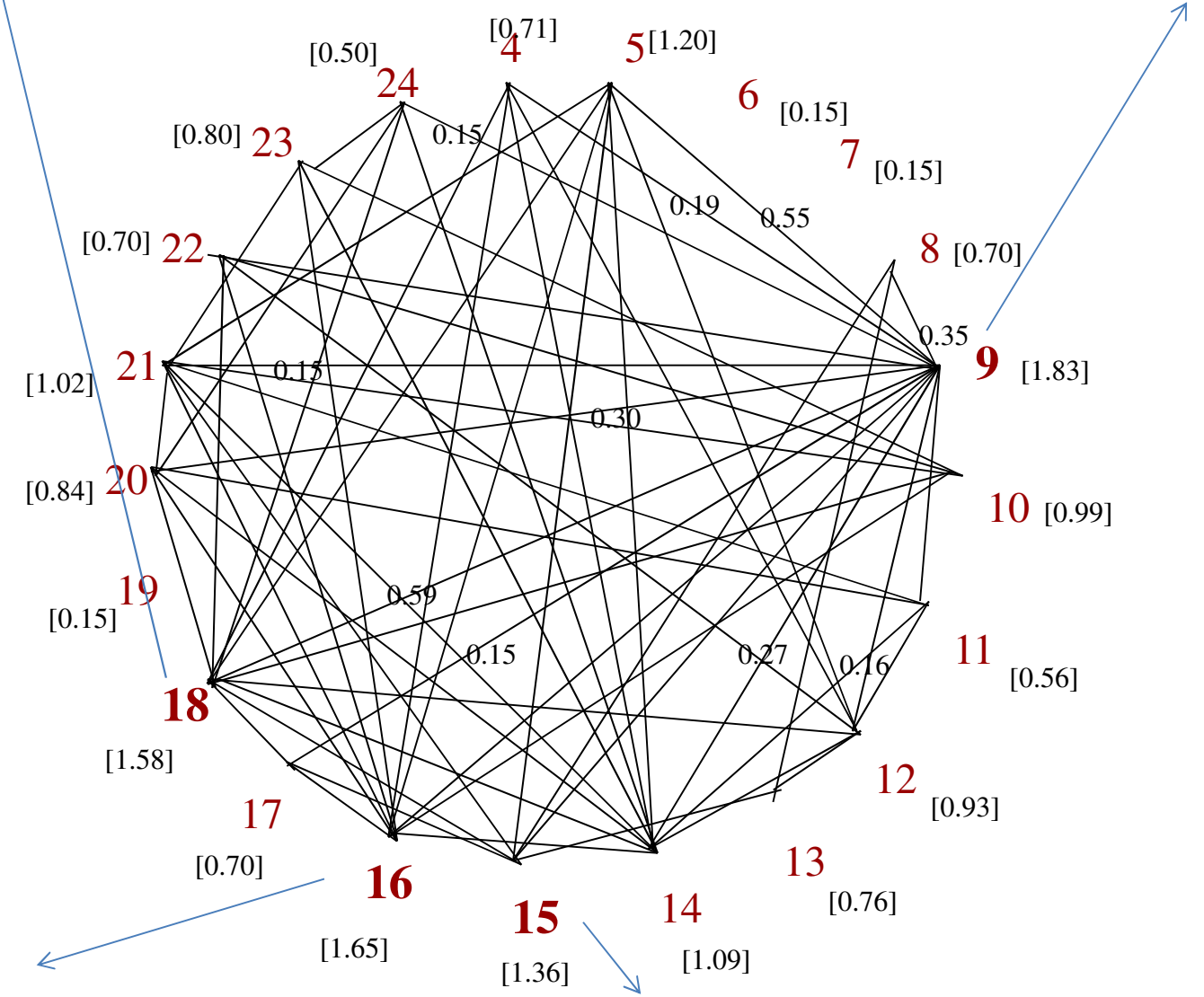
Document summarization overview

Text about “Hurricane Hilbert” (24 sentences) - TextRank

3. r i BC-HurricaneGilbert 09-11 0339
4. BC-Hurricane Gilbert , 0348
5. Hurricane Gilbert Heads Toward Dominican Coast
6. By RUDDY GONZALEZ
7. Associated Press Writer
8. SANTO DOMINGO , Dominican Republic (AP)
9. Hurricane Gilbert swept toward the Dominican Republic Sunday , and the Civil Defense alerted its heavily populated south coast to prepare for high winds , heavy rains and high seas .
10. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph .
11. " There is no need for alarm , " Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday .
12. Cabral said residents of the province of Barahona should closely follow Gilbert 's movement .
13. An estimated 100,000 people live in the province , including 70,000 in the city of Barahona , about 125 miles west of Santo Domingo .
14. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night
15. The National Hurricane Center in Miami reported its position at 2a.m. Sunday at latitude 16.1 north , longitude 67.5 west , about 140 miles south of Ponce , Puerto Rico , and 200 miles southeast of Santo Domingo .
16. The National Weather Service in San Juan , Puerto Rico , said Gilbert was moving westward at 15 mph with a " broad area of cloudiness and heavy weather " rotating around the center of the storm .
17. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6p.m. Sunday .
18. Strong winds associated with the Gilbert brought coastal flooding , strong southeast winds and up to 12 feet to Puerto Rico 's south coast .
19. There were no reports of casualties .
20. San Juan , on the north coast , had heavy rains and gusts Saturday , but they subsided during the night .
21. On Saturday , Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast .
22. Residents returned home , happy to find little damage from 80 mph winds and sheets of rain .
23. Florence , the sixth named storm of the 1988 Atlantic storm season , was the second hurricane .
24. The first , Debby , reached minimal hurricane strength briefly before hitting the Mexican coast last month

Strong winds associated with the Gilbert brought coastal flooding , strong southeast winds and up to 12 feet to Puerto Rico 's south coast .

Hurricane Gilbert swept toward the Dominican Republic Sunda and the Civil Defense alerted its heavily



The National Hurricane Center in Miami reported its position at 2a.m. Sunday at latitude 16.1 north , longitude 67.5 west , about 140 miles south of Ponce , Puerto Rico , and 200 miles southeast of Santo Domingo .

The National Weather Service in San Juan , Puerto Rico , said Gilbert was moving westward at 15 mph with a " broad area of cloudiness and heavy weather " rotating around the center of the storm .

Document summarization overview

Text about “Hurricane Hilbert” (24 sentences) – TextRank Summary

- Automatic summary

  Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The National Hurricane Center in Miami reported its position at 2a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a " broad area of cloudiness and heavy weather " rotating around the center of the storm. Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet to Puerto Rico's coast.

- Reference summary I

  Hurricane Gilbert swept toward the Dominican Republic Sunday with sustained winds of 75 mph gusting to 92 mph. Civil Defense Director Eugenio Cabral alerted the country's heavily populated south coast and cautioned that even though there is no nee d for alarm, residents should closely follow Gilbert's movements. The U.S. Weather Service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday. Gilbert brought coastal flooding to Puerto Rico's south coast on Saturday. There have been no reports of casualties. Meanwhile, Hurricane Florence, the second hurricane of this storm season, was downgraded to a tropical storm.

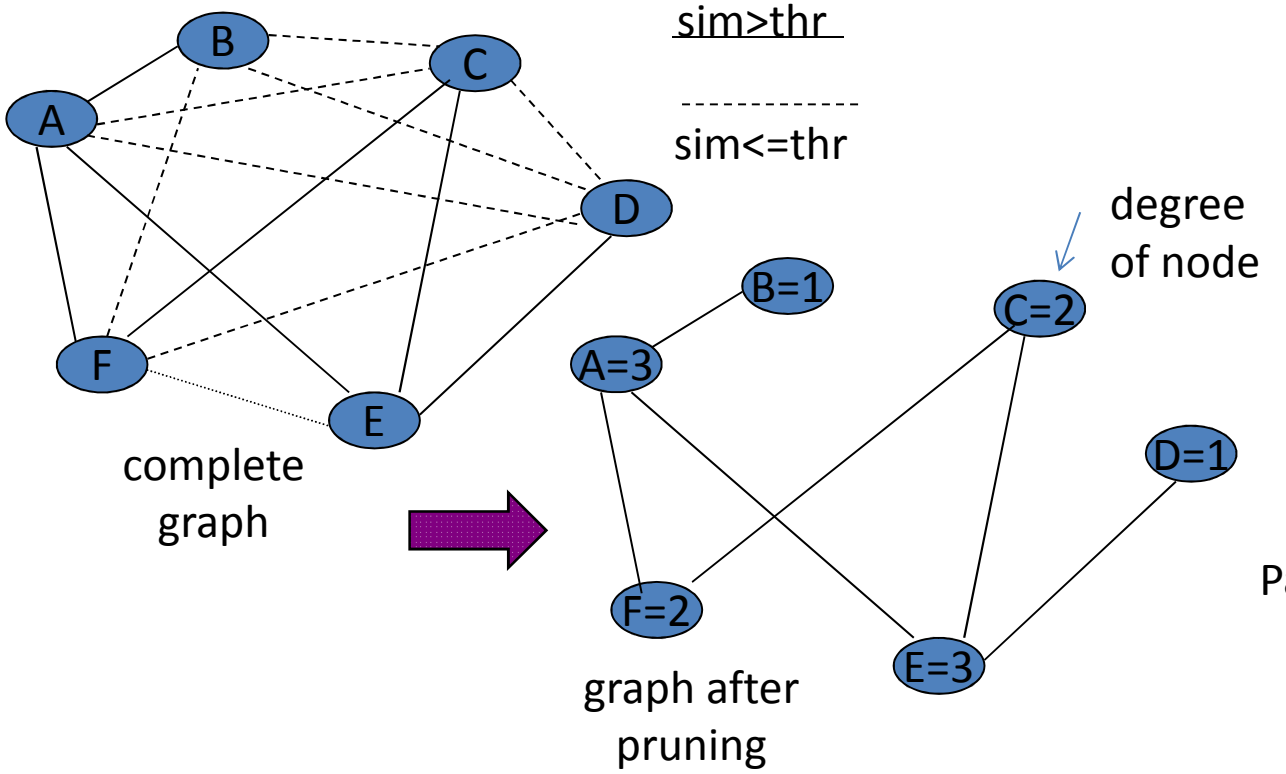
- Reference summary II

  Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, hav e been alerted to prepare for heavy rains, and high winds and seas. Tropical Storm Gilbert formed in the eastern Caribbean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical storm.

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



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

$$sim(D_i, D_j) = \sum d_{ik} . d_{jk}$$

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

(Salton et al. 1997)

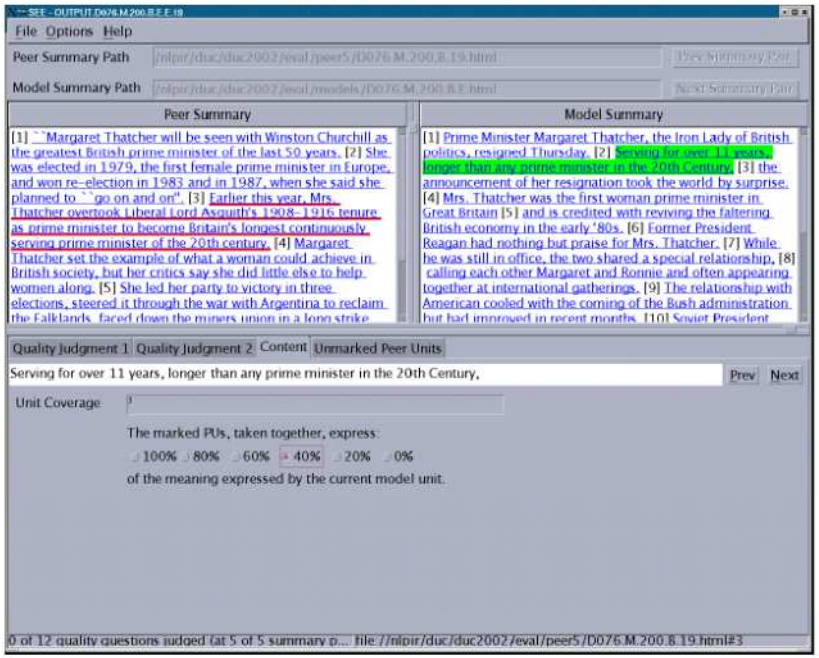
Document summarization overview

Evaluation of automated summaries

Human assessment of content: **check with source document or with ideal summaries**, and **text quality**: grammaticality, coherence, etc.

- 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 INTERFACE



DUC evaluations

Document summarization overview

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

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

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)

Document summarization overview

Evaluation of automated summaries

Computing ROUGE-1 and ROUGE-L

- **Summary:** At least 13 sailors have been killed in a mine attack on a convoy in north-western Sri Lanka, officials say.
- **Model-1:** Tamil Tiger guerrillas have blown up a navy bus in northeastern Sri Lanka, killing at least 10 sailors and wounding 17 others.
- **Model-2:** Blasts blamed on Tamil Tiger rebels killed 13 people on Wednesday in Sri Lanka's northeast and dozens more were injured, officials said, raising fears planned peace talks may be cancelled and a civil war could restart.

ROUGE-1

- Peer has 21 1-grams (x2 = 42)
- Model-1 has 22
- Model-2 has 37 (total = 59)
- 1-grams hits 16
- 1-gram recall 0.27
- 1-gram precision 0.38
- 1-gram f-score 0.31

ROUGE-L

LCS: have a in sri lanka

LCS: killed on in sri lanka officials

- Peer has 21 words (x2 = 42)
- Model-1 has 22
- Model-2 has 37 (total = 59)
- LCS-hits is 11
- LCS recall 0.18
- LCS precision 0.26
- LCS f-score 0.21

Document summarization overview

Evaluation of automated summaries

Pyramid Score

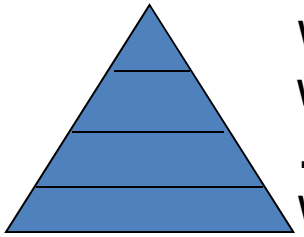
based on the distribution of content units (Sus) in a set of ideal summaries, similar content units are grouped together

- 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:

$$D = \sum_{i=1}^{n} i * D_i$$

- X is the number of units in the summary

$$Score = D / Max$$



PYRAMID FROM CONTENT UNITS IN IDEAL SUMMARIES

$$Max = \sum_{i=j+1}^{n} i * |T_i| + j * (X - \sum_{i=j+1}^{n} |T_i|)$$

$$j = \max_i (\sum_{t=i}^{n} |T_t| \geq X)$$

(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 http://www.summarization.com/mead/
  - 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/
  - 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)

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

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"

Summarizing Scientific Articles

Domain Specific Summarization using Information Extraction and Template-based Generation

Summaries in specific scientific domains report information on **specific and stereotypical domain concepts**. The way the information is presented in the summary is also predictable.

- Example in the are 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 G. pallida on the yield of potato. An experiment in 1985 and 1986 at York was undertaken.

(Oakes and Paice, 2000)

Scientific document summarization

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

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)

Scientific document summarization

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

DICTIONARY

- **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*) ➔ impersonal (e.g. *Describes X)* etc.
- **Evaluation**: text classification, comparison against author abstract, comparison against ideal summaries
  - Improves over all baselines

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.

BioSumm 2014 & SciSumm 2016

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

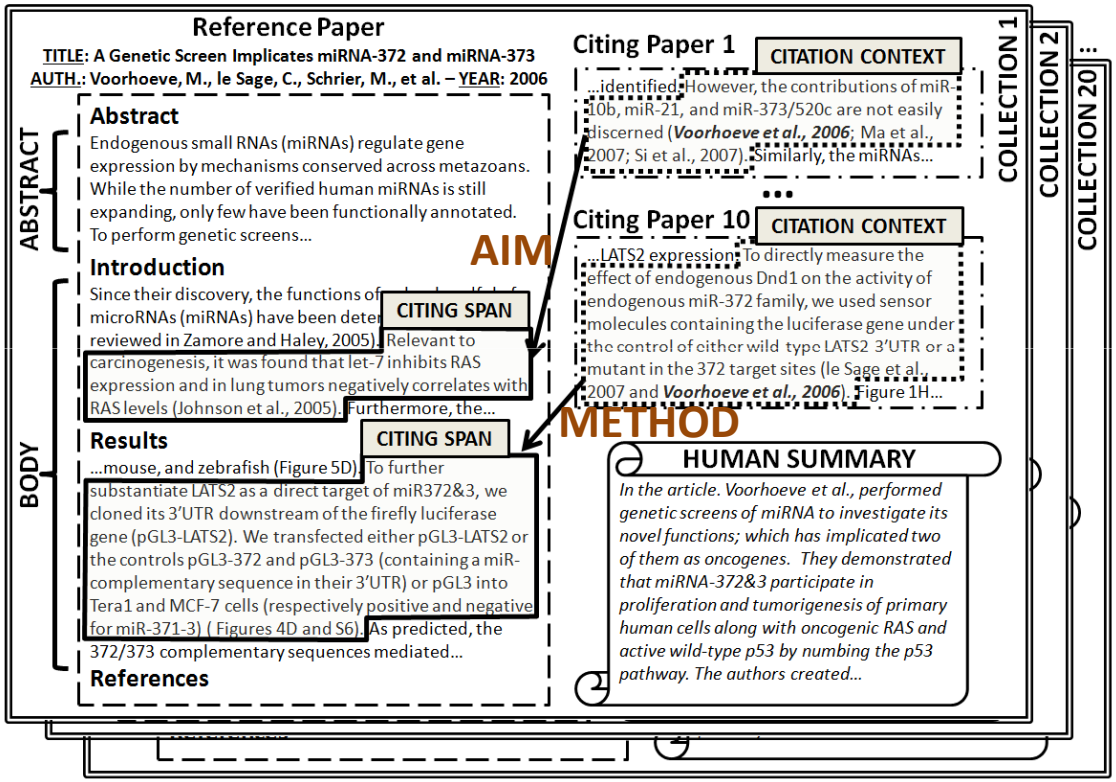
- **Author abstracts** do not provide information on the lasting influences of a work
- **Citations** do not provide enough context from the cited paper

https://tac.nist.gov//2014/BiomedSumm/

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 training and **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

https://tac.nist.gov//2014/BiomedSumm/

http://wing.comp.nus.edu.sg/birndl-jcdl2016/

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)

Summarizing Scientific Articles

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

	Fact	Occurrences
Shared	f_4: "Czech DP"	10
	f_1: "lexical rules"	6
	f_3: "POS/ tag classification"	6
	f_2: "constituency parsing"	5
	f_5: "Punctuation"	2
	f_6: "Reordering Technique"	2
	f_7: "Flat Rules"	2
Unshared	"Dependency conversion"	
	"80% UAS"	
	"97.0% F-measure"	
	"Generative model"	
	'Relabel coordinated phrases"	
	'Projective trees"	
	"Markovization"	

Facts for paper "A Statistical Parser for Czech" with 54 citations

- 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

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)

Summarizing Scientific Articles

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

- 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)

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-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)

(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..."

- **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 *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)

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)

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..."

The scoring function of paper sentences (*score(s)*) is based on Kullback-Leibler (KL) divergence

- **V** is the set of words in the vocabulary
- θ_I is the Impact Language Model
- θ_s is the sentence language model $\longrightarrow p(w|\theta_s) = \frac{c(w,s) + \mu_s * P(w|D)}{|s| + \mu_s}$

$$score(s) = -D(\theta_I \,||\, \theta_s) =$$

$$= \sum_{w \in V} p(w \mid \theta_s) \log(p(w \mid \theta_s)) - \sum_{w \in V} p(w \mid \theta_I) \log(p(w \mid \theta_I))$$

If θ_s and θ_I are very close, the KL-divergence would be small and Score(s) would be high

(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)

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations

Finding sentences that potentially contain useful information about a cited source, but not explicitly cite it – i.e. expanding explicit citations to citation contexts

- 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)

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations

- **Method**:

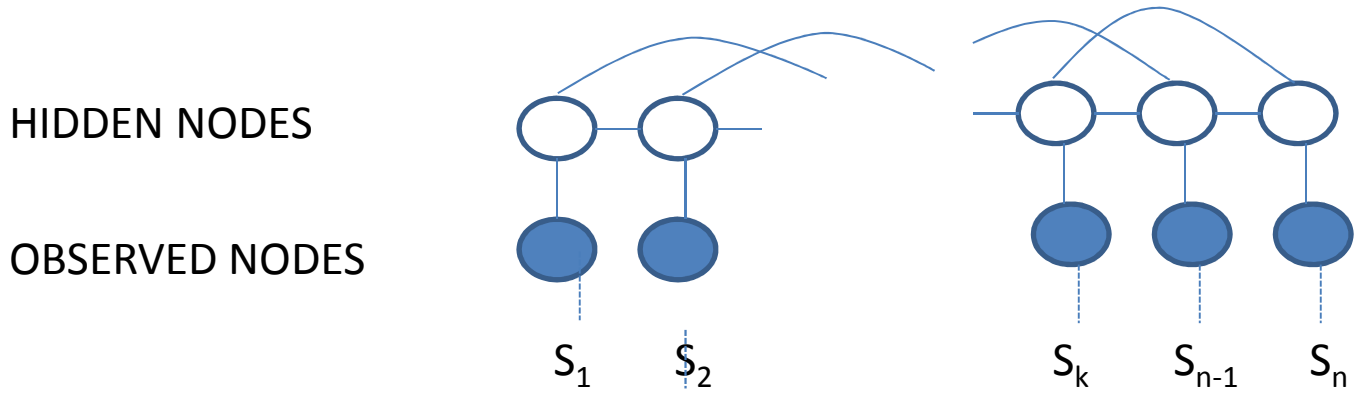
  Construction of a graphical model based on Markov Random Fields (MRF) from the sentences in the document

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

(Qazvinian and Radev, 2010)

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations



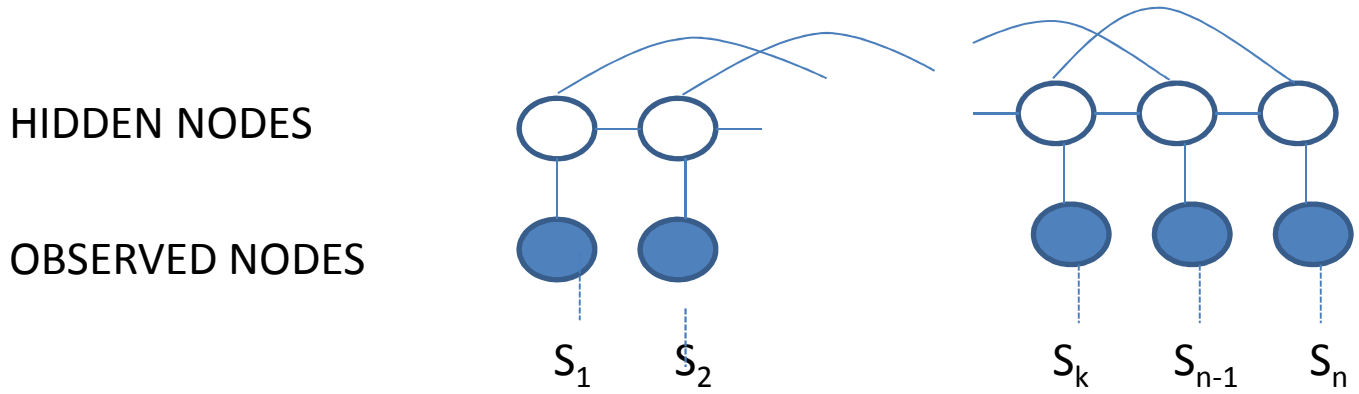
- For each sentence S_i, c_i represents an event of being a non-explicit citation

- Observed nodes represent measurable information about sentences (sentence content)
- Hidden nodes represents the state of the sentence (**non-explicit citation** state) – modelled with a **potential function** $\phi_i(c_i)$ **or probability of being at state** c_i
- Relation between neighbouring sentences represented with a weighted edge: **compatibility function** $\psi_{ij}(c_j \mid c_i)$ represents i believes about j

(Qazvinian and Radev, 2010)

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations



- 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

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

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

(Qazvinian and Radev, 2010)

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations

- 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)*

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

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

- $ne(i)$ 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)

Summarizing Scientific Articles

Citation-based summarization: finding non-explicit citations

- Final *believe values (i.e. probabilities)* are computes with the final values as (with *k* a normalization factor):

$$b(c_i) = k\phi_t(c_i)\prod_{j\in ne(i)} m_{ij}(c_i) \qquad b(\neg c_i) = k\phi_t(\neg c_i)\prod_{j\in ne(i)} m_{ij}(\neg c_i)$$

- Choosing a threshold for deciding if the sentence is a non-explicit citation
- The values of $\phi_t(c_i)$ 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)

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 (BP_1: one neighbour, BP_4: 4 neighbours, BP_n: all neighbours)
- Baseline systems: B_1 selects previous/following sentence if similarity greater than thr. B_2 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)

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

F-score for identifying implicit citations

paper	B_1	B_2	SVM	BP_1	BP_4	BP_n
P08-2026	0.441	0.237	0.249	0.470	**0.613**	0.285
N07-1025	0.388	0.102	0.124	0.313	**0.466**	0.138
N07-3002	0.521	0.339	0.232	**0.742**	0.627	0.315
P06-1101	0.125	0.388	0.127	0.649	**0.889**	0.193
P06-1116	0.283	0.104	0.100	0.307	**0.341**	0.130
W06-2933	0.313	0.100	0.176	0.338	**0.413**	0.160
P05-1044	0.225	0.100	0.060	0.172	**0.586**	0.094
P05-1073	0.144	0.100	0.144	0.433	**0.518**	0.171
N03-1003	0.245	0.249	0.126	**0.523**	0.466	0.125
N03-2016	0.100	0.181	0.224	0.439	**0.482**	0.185

(Qazvinian and Radev, 2010)

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)

Summarizing Scientific Articles

Citation-based summarization: improving coherence

- 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 **B**ackground, **P**roblem, **M**ethod, **R**esult and **L**imitation
  - 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)

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)

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)

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

US007607083B2

(12) **United States Patent**
Gong et al.

(10) **Patent No.: US 7,607,083 B2**
(45) **Date of Patent: Oct. 20, 2009**

(54) **TEST SUMMARIZATION USING RELEVANCE MEASURES AND LATENT SEMANTIC ANALYSIS**

(75) Inventors: **Yihong Gong**, Sunnyvale, CA (US); **Xin Liu**, Berkeley, CA (US)

(73) Assignee: **NEC Corporation**, Tokyo (JP)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 779 days.

(21) Appl. No.: 09/817,591

(22) Filed: **Mar. 26, 2001**

(65) **Prior Publication Data**

US 2002/0138528 A1 Sep. 26, 2002

Related U.S. Application Data

(60) Provisional application No. 60/254,535, filed on Dec. 12, 2000.

(51) **Int. Cl.**
G06F 17/00 (2006.01)

(52) **U.S. Cl.** **715/254**; 705/3; 705/4; 705/5; 707/101; 707/102; 704/245

(58) **Field of Classification Search** 715/530, 715/532, 526, 531; 704/1; 706/45; 707/3, 707/5; 725/116

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,029,195	A *	2/2000	Herz	725/116
6,356,864	B1 *	3/2002	Foltz et al.	704/1
6,523,026	B1 *	2/2003	Gillis	707/3
6,611,825	B1 *	8/2003	Billheimer et al.	706/45
6,865,572	B2 *	3/2005	Boguraev et al.	707/5
2002/0002450	A1 *	1/2002	Nunberg et al.	704/1
2002/0078090	A1 *	6/2002	Hwang et al.	707/513

FOREIGN PATENT DOCUMENTS

JP	2001-014341	A	1/2001

OTHER PUBLICATIONS

Goldstein et al. "Summarizing Text Documents: Sentence Selection and Evaluation Metrics" published Aug. 1999 by ACM Press pp. 121-128.*

(Continued)

Primary Examiner—Doug Hutton
Assistant Examiner—Quoc A Tran
(74) *Attorney, Agent, or Firm*—Sughrue Mion, PLLC

(57) **ABSTRACT**

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 a 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

TEST SUMMARIZATION USING RELEVANCE MEASURES AND LATENT SEMANTIC ANALYSIS

This application claims the benefit of U.S. Provisional Application No. 60/254,535, filed Dec. 12, 2000, entitled "Text Summarization Using IR Technique And Singular Value Decomposition," the disclosure of which is hereby incorporated by reference.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention is related generally to summarization of document contents, and more particularly to a system and method of summarizing the content of text documents through implementation of relevance measurement technologies and latent semantic analysis techniques.

2. Description of the Related Art

The explosive growth of the World-Wide Web has dramatically increased the speed and the scale of information dissemination. With a vast sea of accessible text documents now available on the Internet, conventional Information Retrieval (IR) technologies have become more and more insufficient to find relevant information effectively. Recently, it has become quite common that a keyword-based search on the Internet returns hundreds (or even thousands) of hits, by which the user is often overwhelmed. There is an increasing need for new technologies which assist users in sifting through vast volumes of information, and which can quickly identify the most relevant documents.

Given a large volume of text documents, presenting the user with summaries of these documents greatly facilitates the task of finding documents containing desired information. Text summarization [illegible] nologies that complement each other. Conventional text search engines return a set of documents based upon a relevance measurement with respect to a keyword query, for example; text summarization systems may then produce document summaries that facilitate a quick examination of the contents of each text document returned by the search (by providing, for example, an overview, keyword summary, or abstract).

In other words, a text search engine may typically serve as an information filter for identifying an initial set of relevant documents, while a cooperating text summarization system may serve as an information spotter for assisting the user in identifying a final set of desired or relevant documents.

There are two types of text summaries: generic summaries, and query-relevant summaries. Generic summaries provide an overall sense of a particular document's content, while query-relevant summaries present only content from a particular document that is closely related to the initial search query.

A good generic summary should contain the main topics presented in a document while minimizing redundancy. Since the generic summarization process is not responsive to a particular keyword query 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 document contents that are specifically related to an initial search query; in many existing systems, creating a query-relevant summary is essentially a process of retrieving query-relevant 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, query-relevant summarization is most often achieved simply by extending conventional IR technologies.

Many text summarization methods have been proposed; many recent research studies have been directed toward query-relevant text summarization methods. For example, B. Baldwin and T. S. Morton have proposed a query-sensitive summarization method that selects sentences from the documents until all the phrases in the query are represented. A sentence in the document is considered to represent a phrase in the query if the sentence and the phrase "co-refer" to the same individual, organization, event, and so forth (B. Baldwin et al., *Dynamic Co-reference-Based Summarization*, in Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP3), Granada, Spain, June 1998). R. Barzilay and M. Elhadad have developed a method that creates text summaries by finding lexical chains in documents (R. Barzilay et al., *Using Lexical Chains For Text Summarization*, in Proceedings of the Workshop on Intelligent Scalable Text Summarization (Madrid, Spain), August 1997).

Mark Sanderson has approached the problem by dividing each document into equally sized overlapping passages, and using the INQUERY IR system to retrieve the passage from each document that best matches a query. This "best passage" is then used as a summary of the document. A query expansion technique called Local Context Analysis (LCA, which is also from INQUERY) is used before the best passage retrieval. Given a topic and a document collection, the LCA procedure retrieves top-ranked documents from the collection and examines the context surrounding the topic terms in each retrieved document; LCA then selects the words or phrases that are frequent in these contexts and adds these words or phrases to the original query (M. Sanderson, *Accurate User Directed Summarization From Existing Tools*, in Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98), 1998).

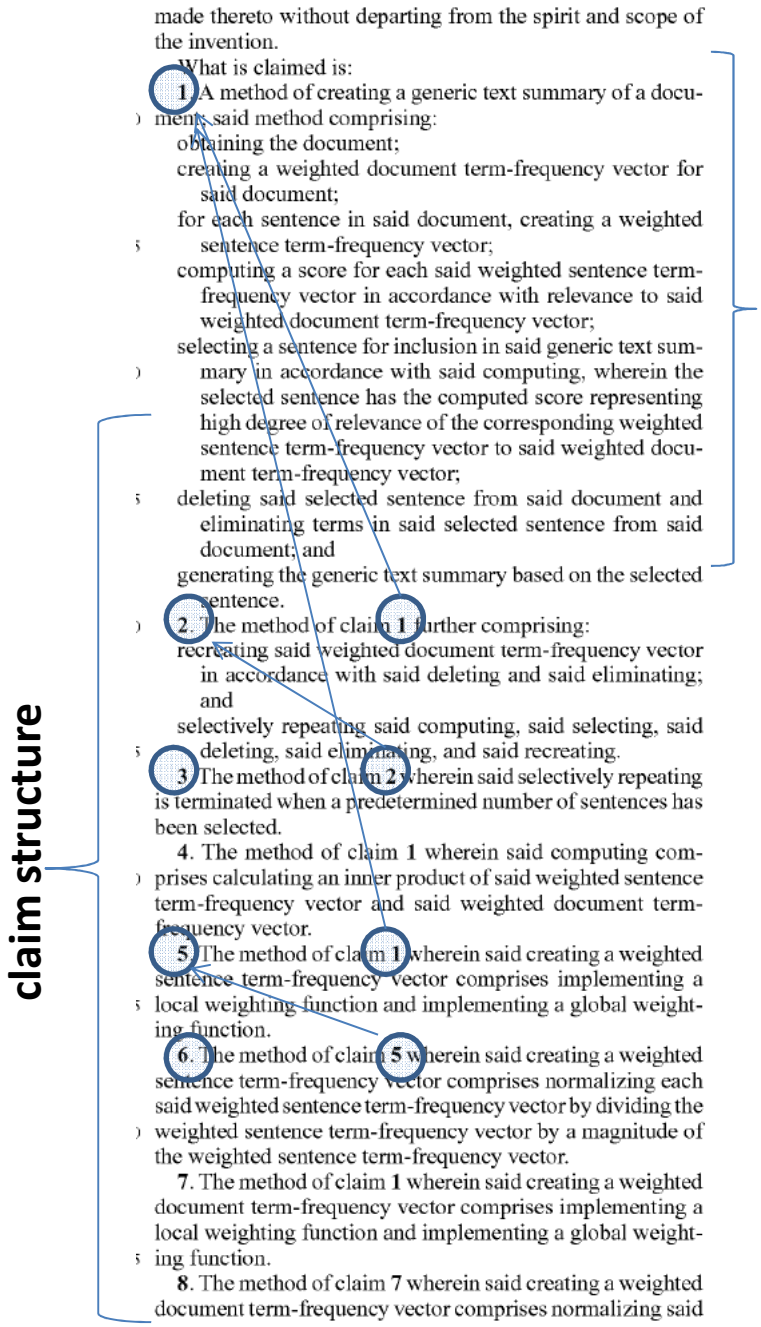
The SUMMARIST text summarizer from the University of Southern California attempts to create text summaries based on the equation:

summarization=topic identification+interpretation+generation

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 portions of the input, or by creating new sentences based on the interpretation of the document concepts (E. Hovy et al., *Automated Text Summarization in Summarist*, in Proceedings of the TIPSTER Workshop, Baltimore, Md., 1998). This generation function 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 combination with statistical information from the corpus to identify the best sentences to include in a summary (http://www.SRA.com). The Cornell/Sabir system uses the document ranking and passage retrieval capabilities of the SMART text search engine to identify relevant passages in a document (C. Buckley et al., *The SMART/Empire TIPSTER IR System*, in Proceedings of TIPSTER Phase III Workshop, 1999). The text summarizer from CGI/CMU uses a technique called Maximal Marginal Relevance (MMR), which mea-

CLAIMS SECTION OF PATENT



made thereto without departing from the spirit and scope of the invention.

What is claimed is:

1. A method of creating a generic text summary of a document, said method comprising:

obtaining the document;

creating a weighted document term-frequency vector for said document;

for each sentence in said document, creating a weighted sentence term-frequency vector;

computing a score for each said weighted sentence term-frequency vector in accordance with relevance to said weighted document term-frequency vector;

selecting a sentence for inclusion in said generic text summary in accordance with said computing, wherein the selected sentence has the computed score representing high degree of relevance of the corresponding weighted sentence term-frequency vector to said weighted document term-frequency vector;

deleting said selected sentence from said document and eliminating terms in said selected sentence from said document; and

generating the generic text summary based on the selected sentence.

2. The method of claim **1** further comprising:

recreating said weighted document term-frequency vector in accordance with said deleting and said eliminating; and

selectively repeating said computing, said selecting, said deleting, said eliminating, and said recreating.

3. The method of claim **2** wherein said selectively repeating is terminated when a predetermined number of sentences has been selected.

4. The method of claim **1** wherein said computing comprises calculating an inner product of said weighted sentence term-frequency vector and said weighted document term-frequency vector.

5. The method of claim **1** wherein said creating a weighted sentence term-frequency vector comprises implementing a local weighting function and implementing a global weighting function.

6. The method of claim **5** wherein said creating a weighted sentence term-frequency vector comprises normalizing each said weighted sentence term-frequency vector by dividing the weighted sentence term-frequency vector by a magnitude of the weighted sentence term-frequency vector.

7. The method of claim **1** wherein said creating a weighted document term-frequency vector comprises implementing a local weighting function and implementing a global weighting function.

8. The method of claim **7** wherein said creating a weighted document term-frequency vector comprises normalizing said

claim structure

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

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

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)

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

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

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

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:

[a device (for) containing a signal processing unit]
[a device to contain a signal processing unit]
[a device which contains a signal processing unit]
$\Rightarrow$ *A device contains a signal processing unit*

[a unit contained in a rectangular device] $\Rightarrow$ *A unit is contained in a rectangular device*

[a sail for a sailboard]$_{invention=initial}$
[a device for coupling a sail batten to a mast in a board sail]$_{invention=\neg initial}$
[a first end for rotateably bearing against a mast]$_{component=yes}$
$\Rightarrow$ *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 ➔ 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

Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '98).

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.* 4, 2, Article 4 (May 2007).

Barzilay, R. and Elhadad, M. 1997. Using lexical chains for text summarization. Advances in Automatic Text Summarization 1999.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 427-435.

Daniel C. Marcu. 1998. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. Dissertation. University of Toronto, Toronto, Ont., Canada, Canada.

Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text structuring and summarization. *Inf. Process. Manage.* 33, 2 (March 1997), 193-207.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.* 22, 1 (December 2004), 457-479.

H. P. Edmundson. 1969. New Methods in Automatic Extracting. *J. ACM* 16, 2 (April 1969), 264-285.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.* 2, 2 (April 1958), 159-165.

Horacio Saggion and Guy Lapalme. 2002. Generating indicative-informative summaries with sumUM. *Comput. Linguist.* 28, 4 (December 2002), 497-526.

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)

Cited works (2/3)

Joan Codina-Filbà, Nadjet Bouayad-Agha, Alicia Burga, Gerard Casamayor, Simon Mille, Andreas Müller, Horacio Saggion, Leo Wanner, Using genre-specific features for patent summaries, Information Processing & Management, Volume 53, Issue 1, January 2017, Pages 151-174, ISSN 0306-4573

Jones, P.A. and Paice, C.D. A 'Select and Generate' Approach to Automatic Abstracting. 14th Information Retrieval Colloquium. 1993.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '95).

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Inf. Process. Manage.* 43, 6 (November 2007), 1449-1481.

Lin, CY. and Hovy, E.. Identifying Topics by Position. ACL 1997.

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.

Nenkova, A. Vanderwende, L. The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101.

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

Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab.

Paice, C.D. Constructing Literarure Abstracts by Computers: Techniques and Prospects. IP&M 1990.

Qiaozhu Mei, ChengXiang Zhai. Generating impact-based summaries for scientific literature. ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference.

Rada Mihalcea, Paul Tarau. 2004. TextRank: Bringing Order into Texts. *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Cited works (3/3)

Rahul Jha, Reed Coke, and Dragomir Radev. 2015. Surveyor: a system for generating coherent survey articles for scientific topics. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (AAAI'15). AAAI Press 2167-2173.

Seiji Miike, Etsuo Itoh, Kenji Ono, and Kazuo Sumita. 1994. A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '94)

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1* (COLING '08), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 689-696.

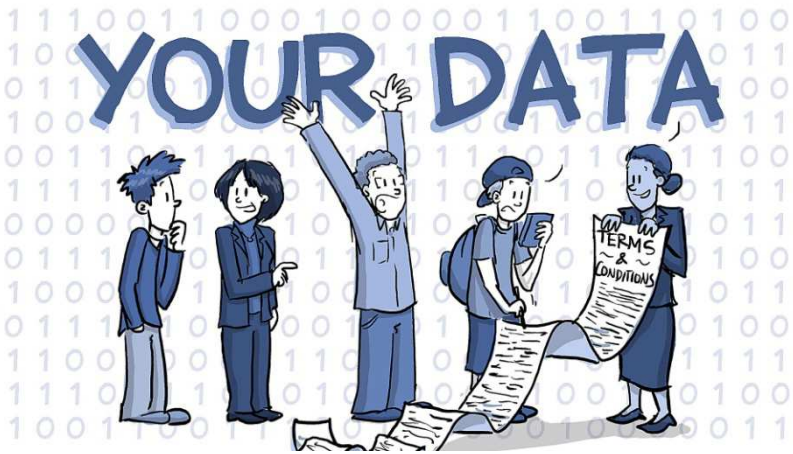
Vahed Qazvinian and Dragomir R. Radev. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 555-564.

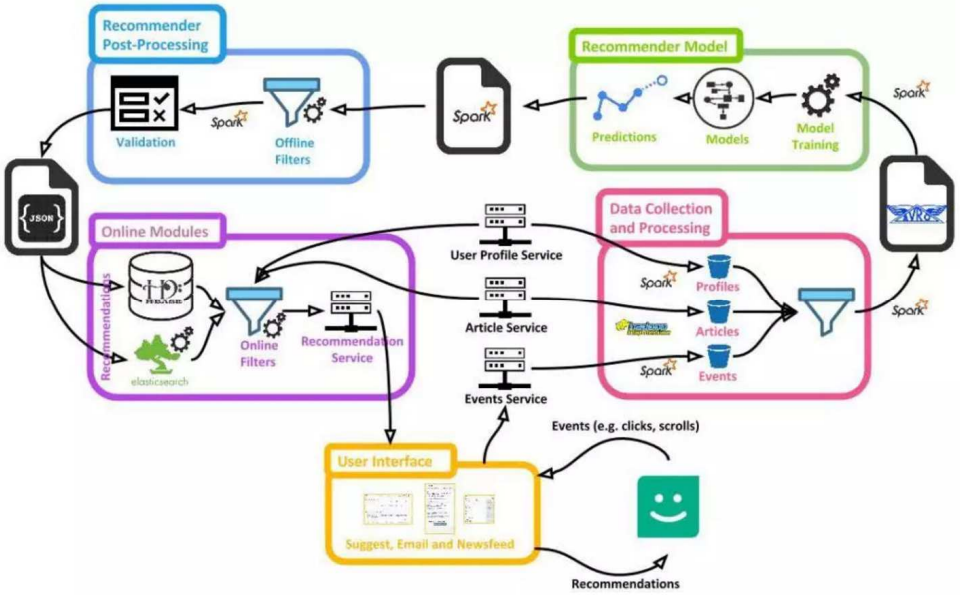
Universitat Pompeu Fabra Barcelona

taln

EXCELENCIA MARÍA DE MAEZTU

CHALLENGES, DATASETS AND ARCHITECTURES





Outline

- Scientific Literature Mining Challenges
- Datasets and tools
- Structured / semantic publication formats
- Scholary literature architectures

Scientific Literature Mining Challenges

Several challenges have been organized
to explore how we can take advantage of scientific literature
to automatically carry out specific text analysis tasks

KDD Cup 2016

KDD Cup 2013

Semantic Publishing Challenge

TAC 2014 Biomedical Summarization Track

CL-SciSumm 2016

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

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

Scientific Literature Mining Challenges

KDD Cup 2013

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

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

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

Roy, S. B., De Cock, M., Mandava, V., Savanna, S., Dalessandro, B., Perlich, C., ... & Hamner, B. (2013, August). *The microsoft academic search dataset and kdd cup 2013.* In Proceedings of the 2013 KDD cup 2013 workshop (p. 1). ACM.

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

Roy, S. B., De Cock, M., Mandava, V., Savanna, S., Dalessandro, B., Perlich, C., ... & Hamner, B. (2013, August). *The microsoft academic search dataset and kdd cup 2013.* In *Proceedings of the 2013 KDD cup 2013 workshop* (p. 1). ACM.

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

Wade, A. D., Wang, K., Sun, Y., Gulli, A. 2016. *WSDM Cup 2016 – Entity Ranking Challenge.* Proceedings of the 9th ACM Conference on Web Search and Data Mining, San Francisco, CA.

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**

Dimou, A., Di Iorio, A., Lange, C., & Vahdati, S. (2016, May). *Semantic Publishing Challenge–Assessing the Quality of Scientific Output in Its Ecosystem.* In Semantic Web Evaluation Challenge (pp. 243-254). Springer International Publishing.

Scientific Literature Mining Challenges

Semantic Publishing Challenge

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

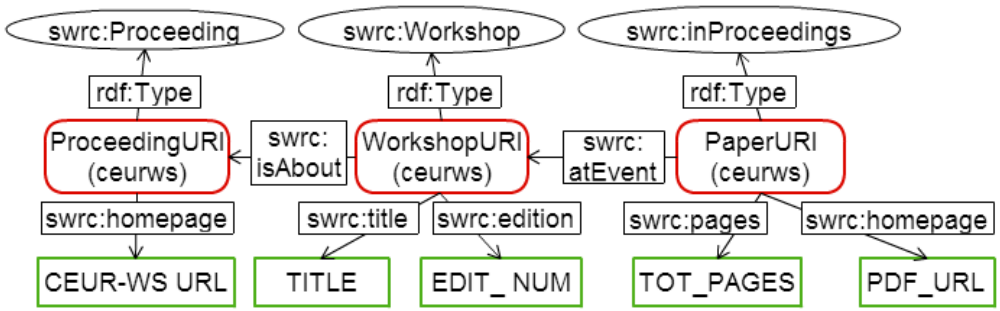
Task 1

Linguisti & structural analyzer → Semantic annotator → Annotation sanitizer → External resources linker → RDF generator → RDF

Task 2

PDF-to-Text converter → Linguistic & structural analyzer → Biblio parser → Spot ontology / founding body mentions → RDF generator → RDF

RDF data model



SPARQL evaluation query

```
SELECT ?procURL (Count( ?paper ) AS ?np)
        (Avg(xsd:int(?numPages)) AS ?al)
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
```

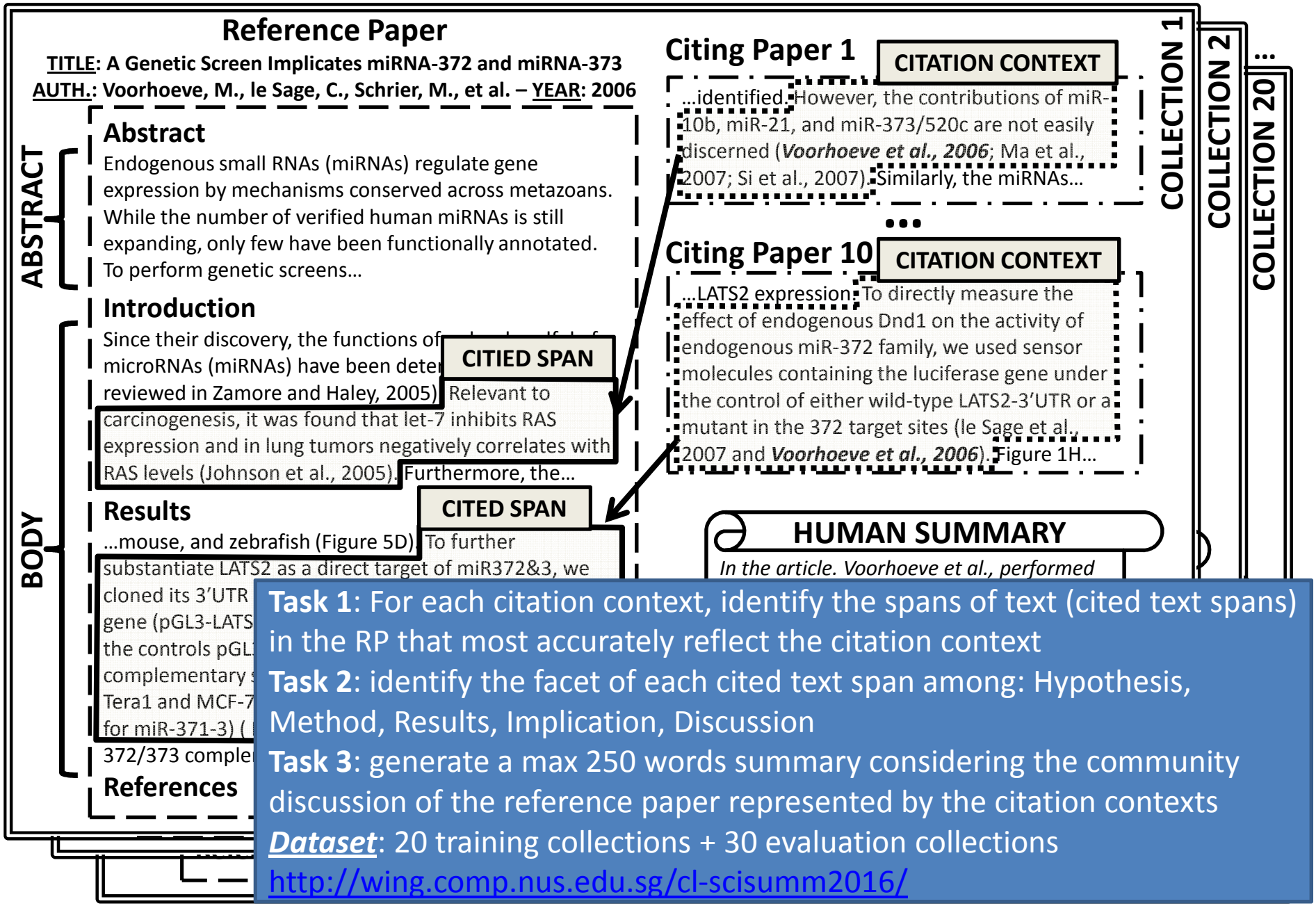
Number and avg. page number of papers in each proceeding

Ronzano, F., Fisas, B., del Bosque, G. C., & Saggion, H. (2015, May). *On the automated generation of scholarly publishing linked datasets: the case of CEUR-WS proceedings.* In Semantic Web Evaluation Challenge (pp. 177-188). Springer International Publishing.

Scientific Literature Mining Challenges

TAC 2014 Biomedical Summarization Track

CL-SciSumm 2016



Task 1: For each citation context, identify the spans of text (cited text spans) in the RP that most accurately reflect the citation context

Task 2: identify the facet of each cited text span among: Hypothesis, Method, Results, Implication, Discussion

Task 3: generate a max 250 words summary considering the community discussion of the reference paper represented by the citation contexts

Dataset: 20 training collections + 30 evaluation collections

http://wing.comp.nus.edu.sg/cl-scisumm2016/

Scientific Literature Mining Challenges

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

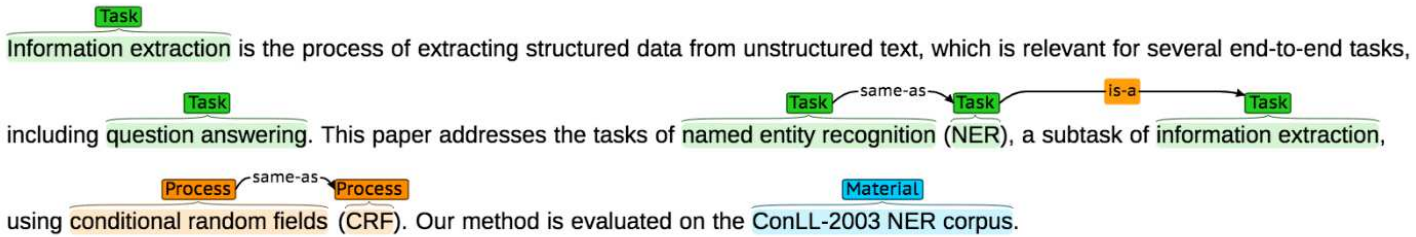
Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. *Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles.* In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 21-26). Association for Computational Linguistics.

- 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

http://alt.qcri.org/semeval2017/task10/ & https://scienceie.github.io/

- 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



Datasets and tools

The ACL anthology network corpus

Radev, D. R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). *The ACL anthology network corpus.* Language Resources and Evaluation, 47(4), 919-944.

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

Number of papers	21,212
Number of authors	17,792
Number of venues	342
Number of paper citations	110,976
Number of author collaborations	142,450
Citation network diameter	22
Collaboration network diameter	15

ACL Anthology SearchBench

Schäfer, U., Kiefer, B., Spurk, C., Steffen, J., & Wang, R. (2011, June). *The ACL anthology searchbench.* In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations (pp. 7-13). Association for Computational Linguistics.

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

Datasets and tools

https://core.ac.uk/

- 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

https://aminer.org/

- 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

CiteSeer^{x} 10M

http://citeseerx.ist.psu.edu/

- 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

https://www.semanticscholar.org/

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

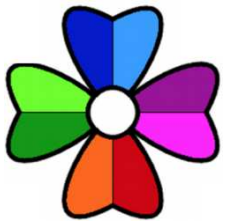
Challenges, datasets and architectures

Structured / semantic publication formats

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



SPAR Ontologies

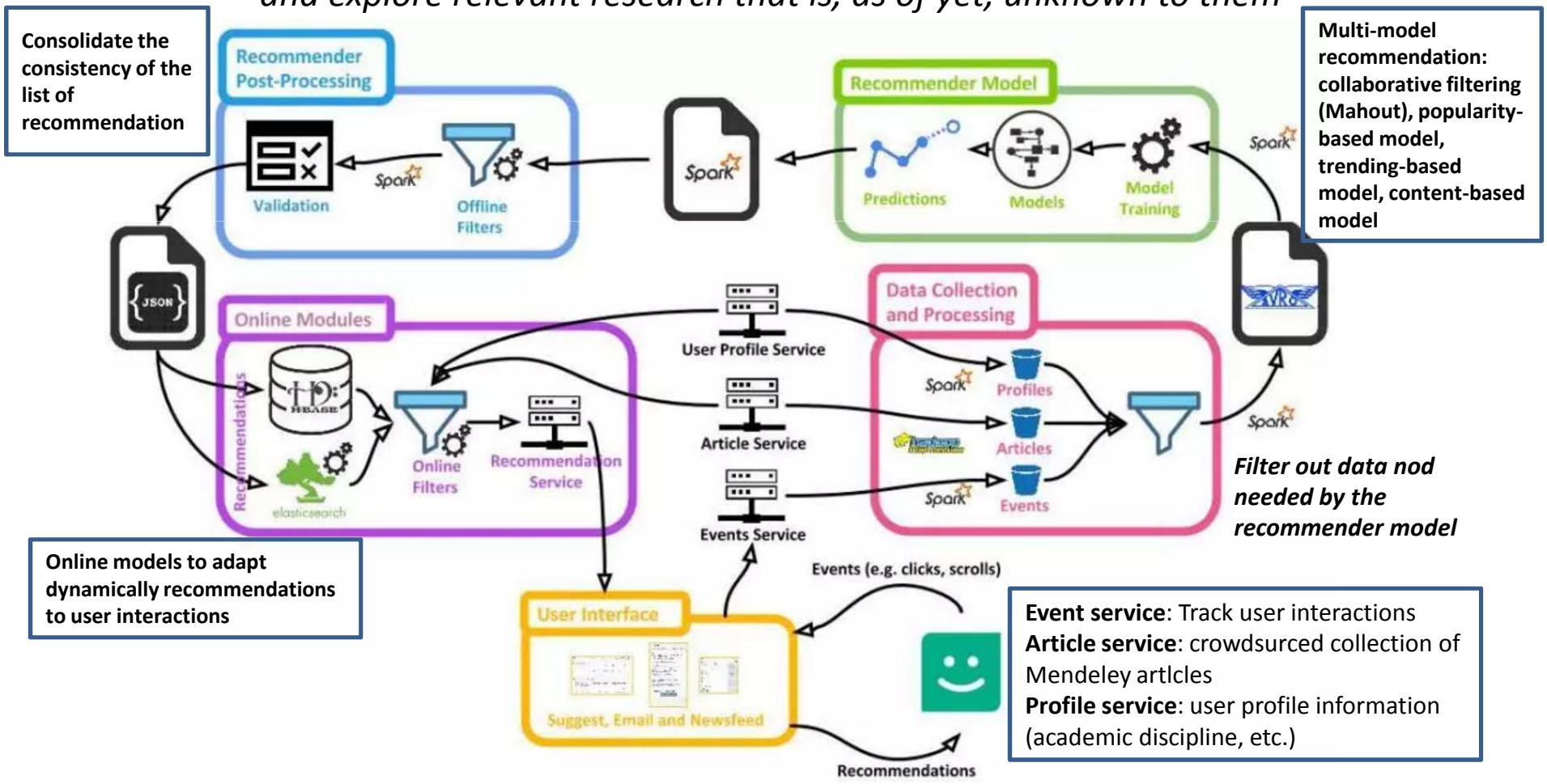
Set of ontologies that support the creation of comprehensive machine-readable RDF metadata for every aspect of semantic publishing and referencing

- 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



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

Crawler: python django

Name disambiguation: RAILS

Universitat Pompeu Fabra Barcelona

EXCELENCIA MARÍA DE MAEZTU

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

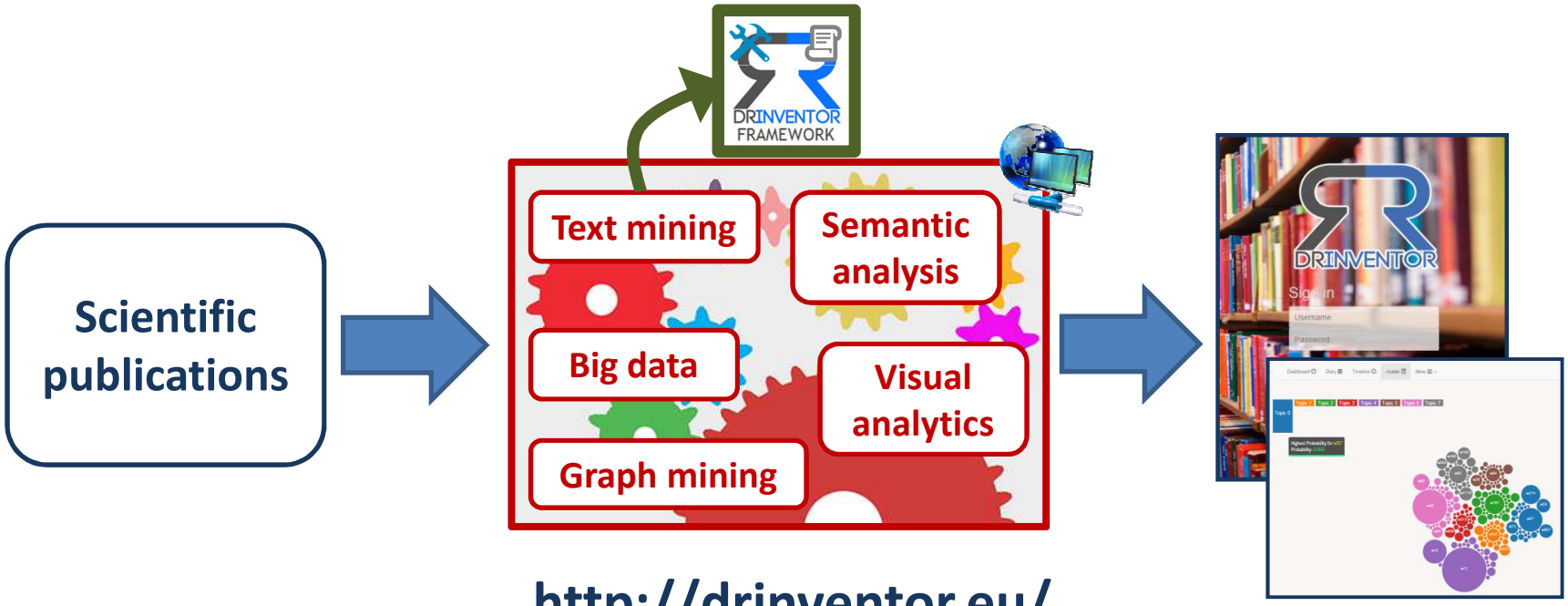
The (bootstrap of) textual analyses of scientific publications often still constitutes a **time-consuming activity** due to:

- ✓ hetherogeneous **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

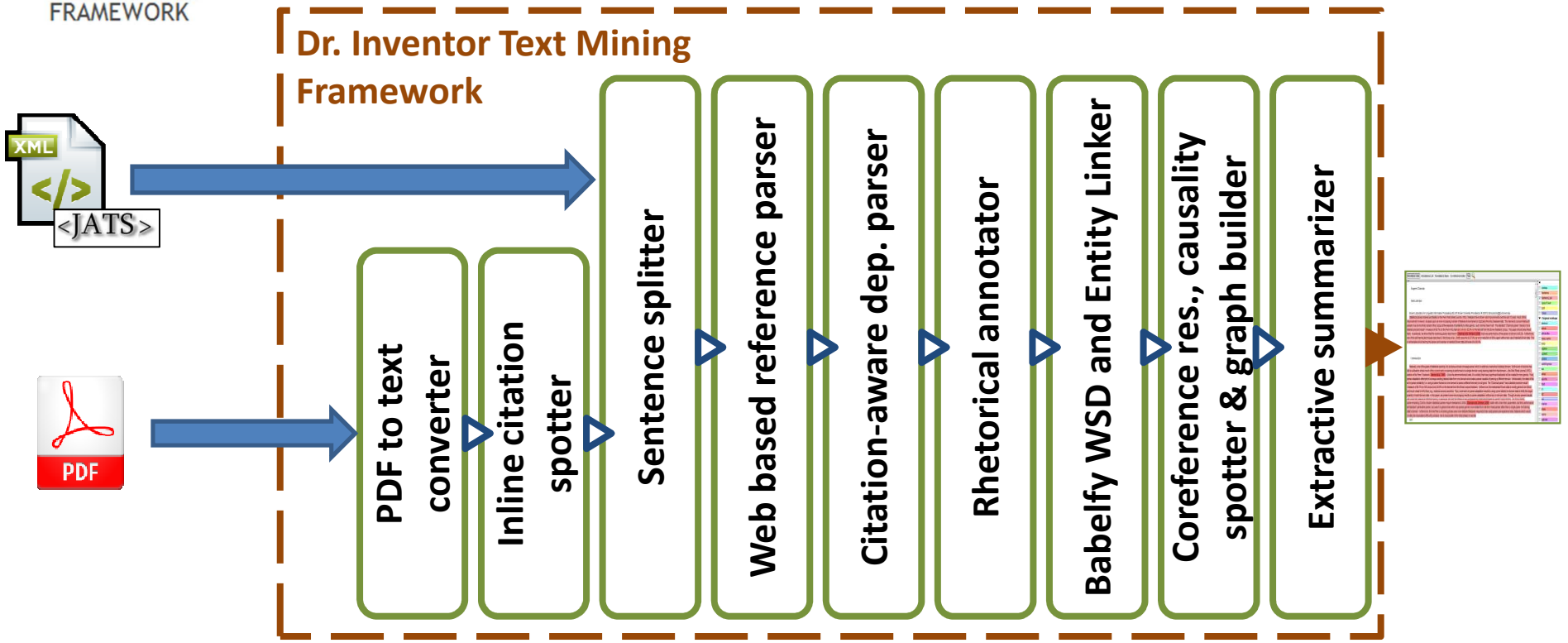
Dr. Inventor Text Mining Framework

- 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 Java8 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 general architecture for text engineering to manage **textual annotations**

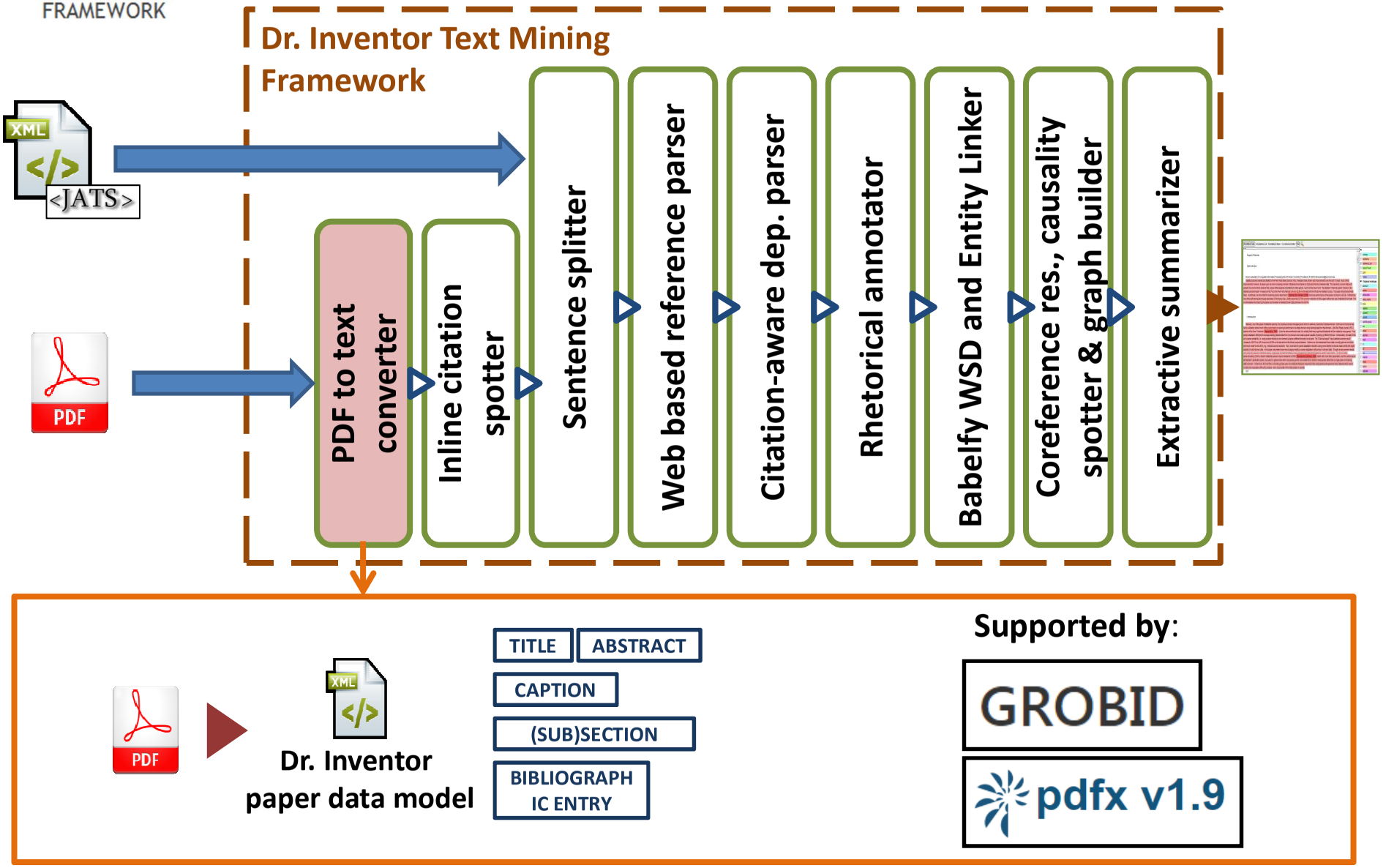
Architectural overview



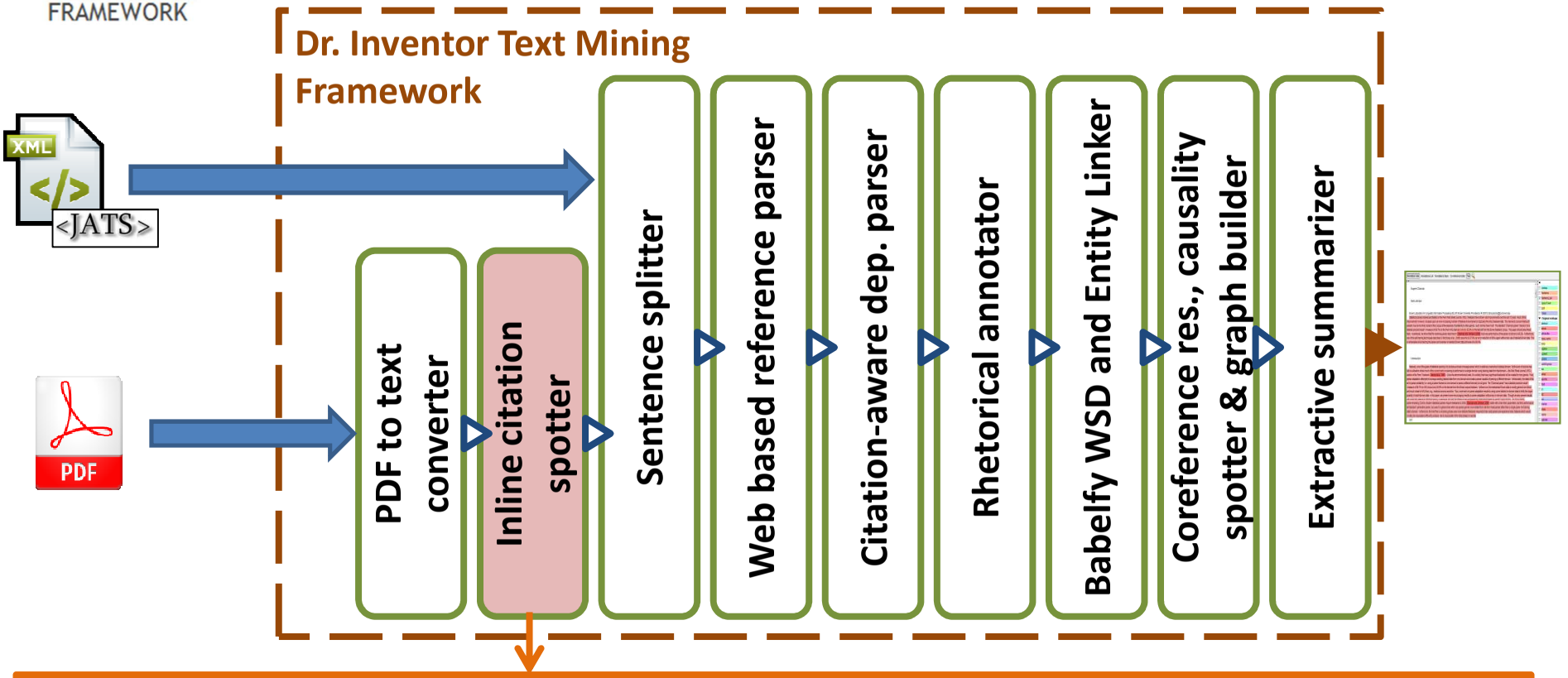
Ronzano, F., & Saggion, H. (2015, October). **Dr. Inventor Framework: Extracting Structured Information from Scientific Publications.** In International Conference on Discovery Science (pp. 209-220). Springer International Publishing.

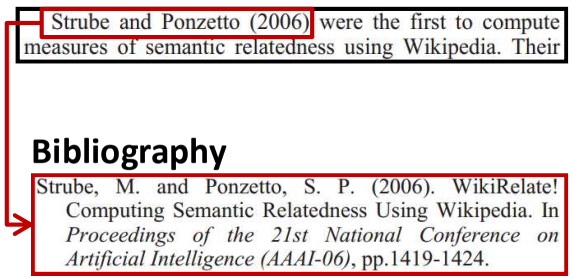
Ronzano, F., & Saggion, H. (2016, April). **Knowledge Extraction and Modeling from Scientific Publications.** In The Semantics, Analytics, Visualization: Enhancing Scholarly Data Workshop, co-located with the 25th International World Wide Web Conference.

Architectural overview



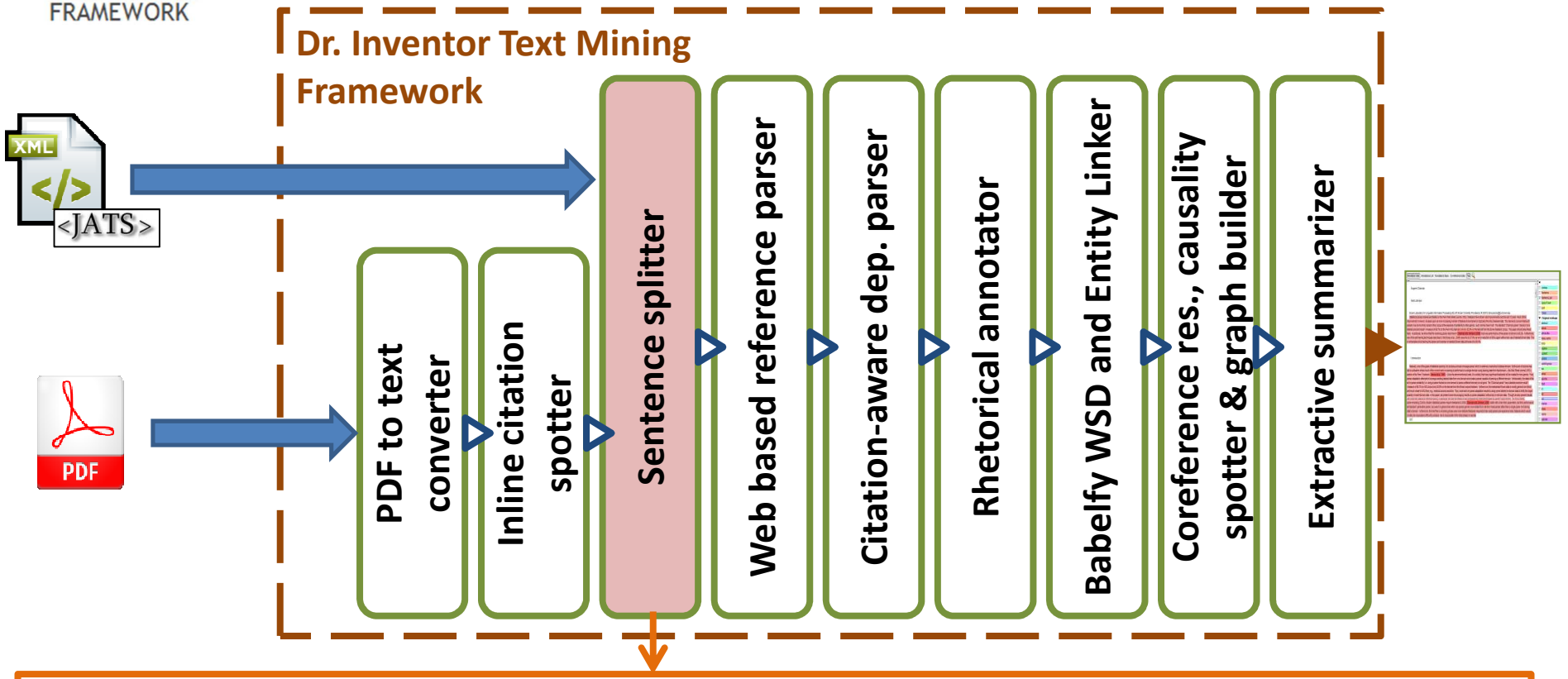
Architectural overview





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

Architectural overview



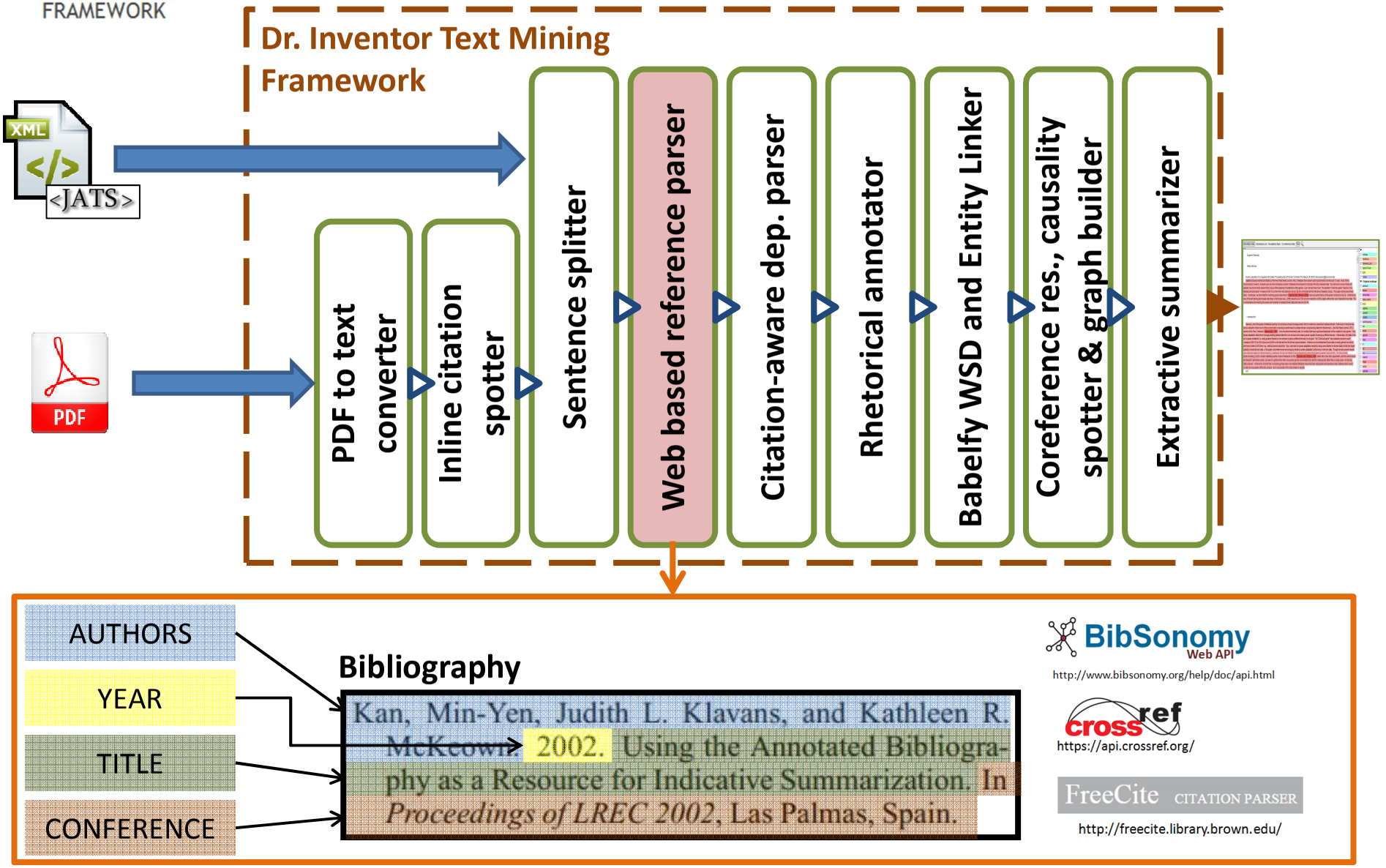
Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. Both (Marcu, 2002) and (Zhang, 2003) consider a sentence pair as different realizations of a sequence of concepts. These alignment approaches segment the sentences into a sequence of phrases.

Customization of ANNIE sentence splitter

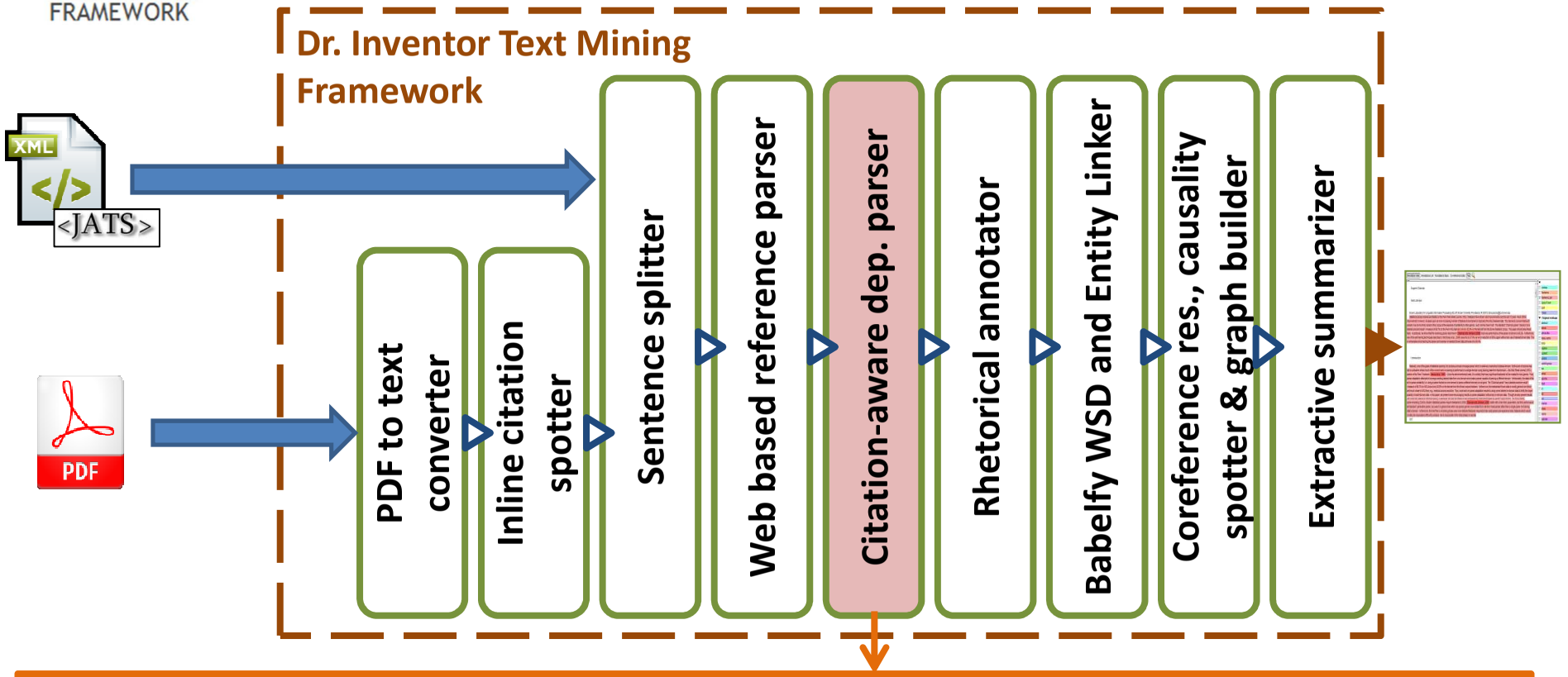
Rule set adapted to peculiarities of scientific papers

by analyzing the **most frequent sentence split patterns / errors** in a set of 40 Computer Graphics papers

Architectural overview

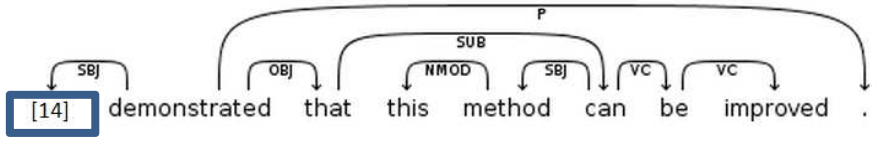


Architectural overview

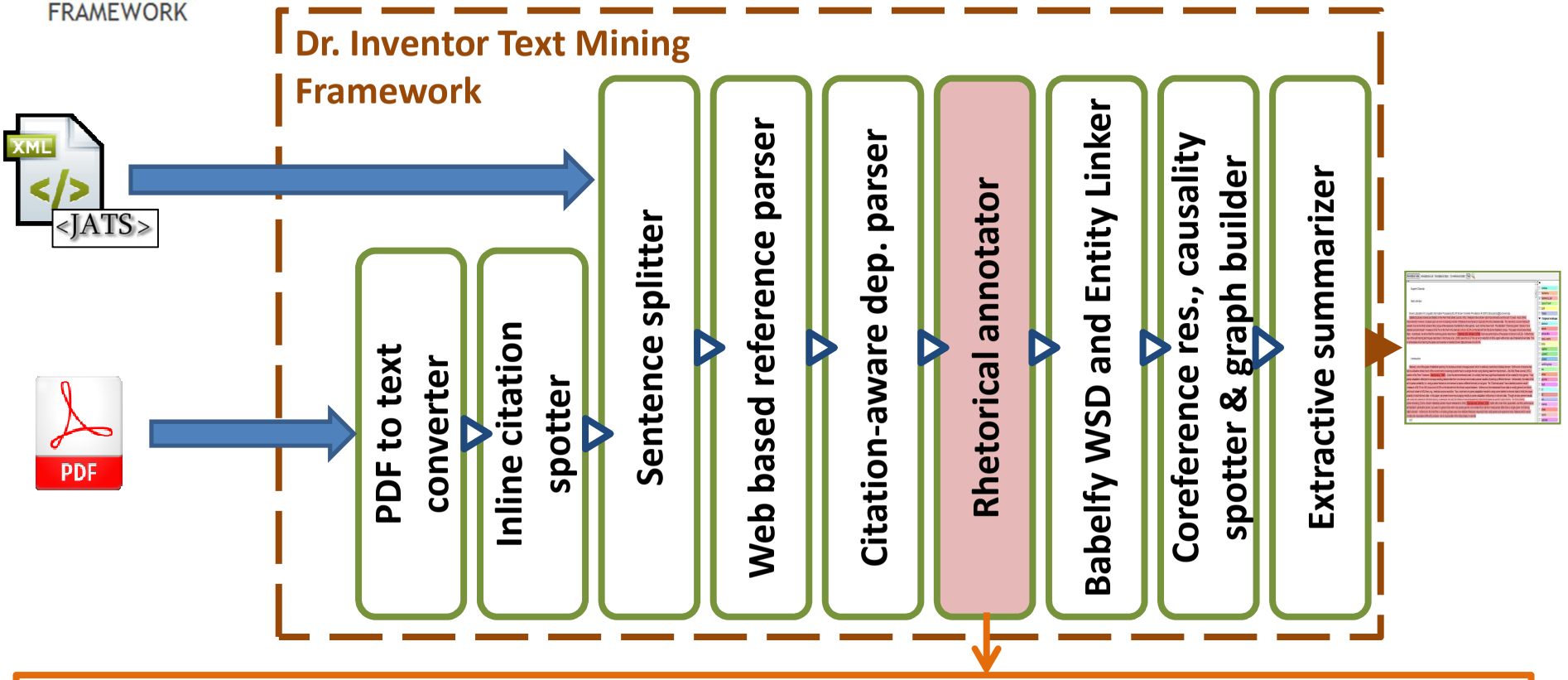


Based on **MATE dependency parser**

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



Architectural overview

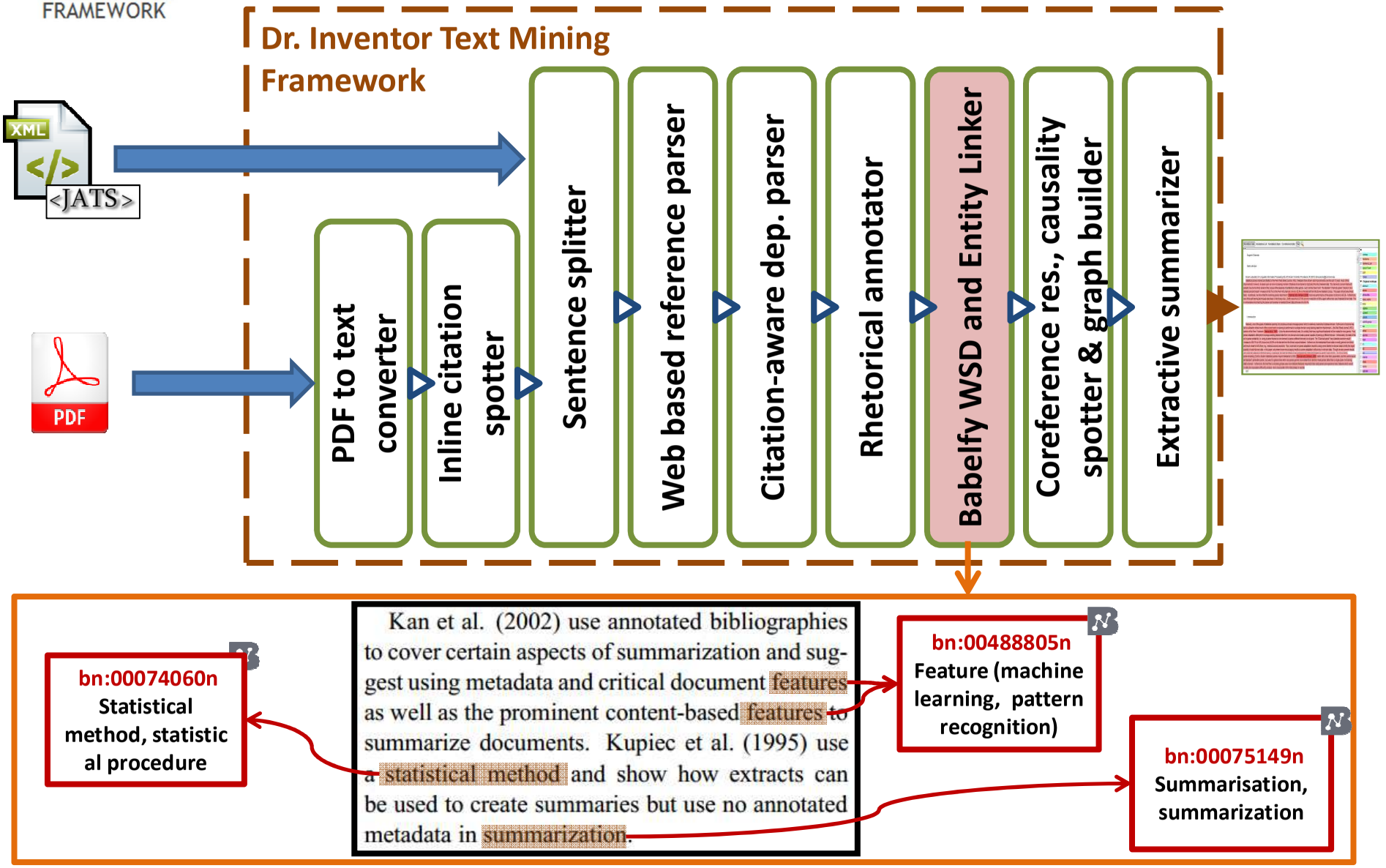


Some alternative phrase alignment approaches have been developed, which do not rely on the Viterbi word alignment. Both (Marcu, 2002) and (Zhang, 2003) consider a sentence pair as different realizations of a sequence of concepts. These alignment approaches segment the sentences into a sequence of phrases.

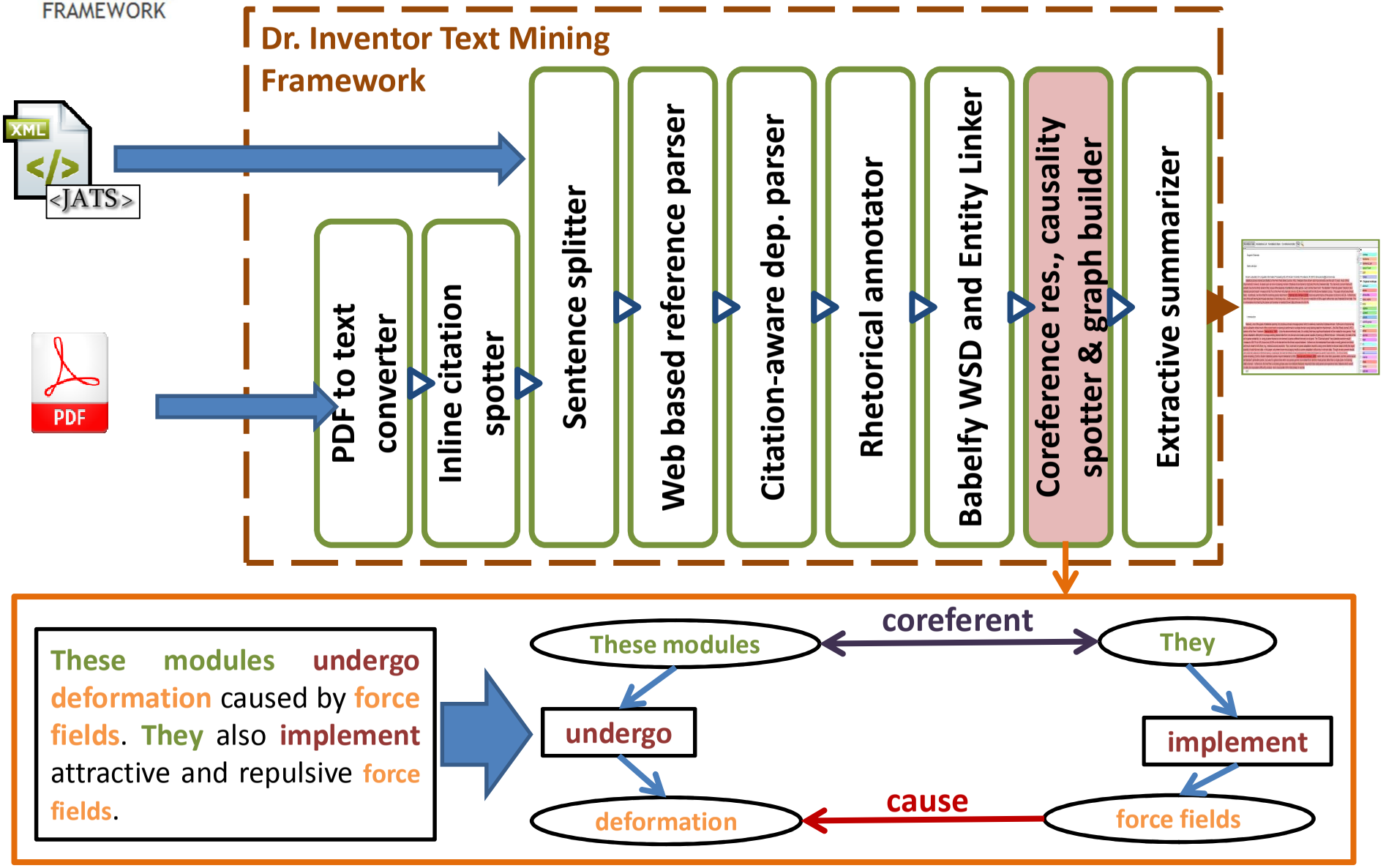
CHALLENGE
BACKGROUND
HYPOTHESIS

- **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)

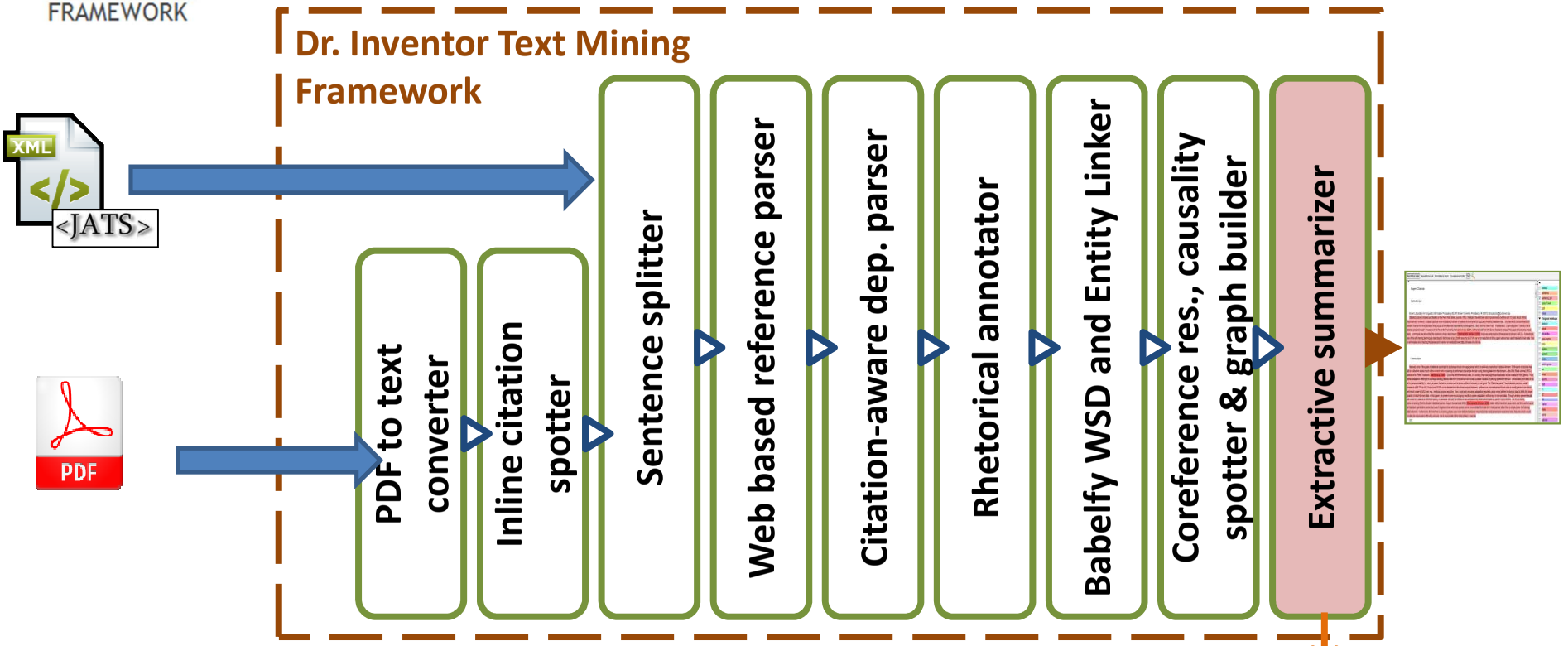
Architectural overview



Architectural overview



Architectural overview



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:

```
<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

Hands-on Dr. Inventor Framework

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

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);
// Improt 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/

Import PDF / JATS XML from file / URL

```
// 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):
Document doc_PDFURL = Factory.getPDFloader().parsePDF(new
URL("http://www2007.org/workshops/paper_45.pdf"));
```

Get ordered lists of sentences

```
// 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);
```

Hands-on Dr. Inventor Framework

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

Print the content of the first sentence of the abstract (by the asString method of the Sentence object instance)

```
// 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

[BABELNET SYNSET] ID: '22047', Text: 'Puppetry', In-sentence ID: '22047', Babel URL: 'http://babelnet.org/rdf/s00065258n', Synset ID: 'bn:00065258n', DBpedia URL: 'http://dbpedia.org/resource/Puppetry', Global score: '2.958149116792834E-4', Coherence score: '0.09968354430379747', Score: '0.8340262582056893', Source: 'bn:00065258n', Num tokens: '1'

[BABELNET SYNSET] ID: '22047', Text: 'art', In-sentence ID: '22047', Babel URL: 'http://babelnet.org/rdf/s00005927n', Synset ID: 'bn:00005927n', DBpedia URL: 'http://dbpedia.org/resource/Art', Global score: '0.023262002991754745', …

Hands-on Dr. Inventor Framework

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

Print the content of the paper sections in document order

```
// 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: '21454', Name: ' 2.1. Head Modelling', Level: '2', Children sections IDs: '[]', Sentences IDs: '[22075, 22064, 22065, 22066, 22067, 22068, 22069, 22070, 22071, 22072, 22073, 22074]'

[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

```
// 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));
}
```

[CITATION] ID: '21440', Source: '[Bibsonomy]', Title: 'Realtime performance-based facial animation.', Year: '2011', Pages: '77',

Bibsonomy URL: 'http://dblp.uni-trier.de/db/journals/tog/tog30.html#WeiseBLP11', Volume: '30', Journal: 'ACM Trans. Graph.',

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: ', Sofien Bouaziz', First name: ', Sofien', Surname: 'Bouaziz'

[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'

Hands-on Dr. Inventor Framework

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

Get an print the list of sentences of the 10-sentences extractive summary generated by the section title tf-idf similarity method

```
// Get ordered list of summary sentences
List<Sentence> summarySentences_TITLE_10 =
doc_PDFfile.extractSummary(20, SummaryTypeENUM.TITLE_SIM);

// Print the text of each sentence
for(Sentence sent : summarySentences_CENTROID_20){
    System.out.println(sent.getText());
}
```

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.

Hands-on Dr. Inventor Framework

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

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 extracter
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**

Scholarly Data Mining: Making Sense of Scientific Literature

THANKS!

Horacio Saggion & Francesco Ronzano

Natural Language Processing Group (TALN)

Universitat Pompeu Fabra, Barcelona, Spain

Tutorial @ JCDL 2017

19th June 2017