

Keywords, phrases, clauses and sentences

Topicality, indicativeness and informativeness at scales

Min-Yen Kan

Web IR / NLP Group (WING)
National University of Singapore
Slides available at: dwz.cn/kan-kp



NUS
National University
of Singapore

School of
Computing

Slides available at: dwz.cn/kan-kp

SOURCES FOR KEYPHRASE EVIDENCE



Motic has launched its new upright microscope, the BA410, a newly designed, modular stand especially for routine-clinical, lab, and teaching applications suitable for a wide range of transmitted light applications for the life science markets.

A completely redesigned optical system ensures that the BA410 will provide the best image quality in the demanding cytology, pathology, and histology fields from both demanding amateur to professional levels. A variety of new viewing heads are also available, including a Trinocular head with three light splits (100:0/20:80/0:100) and two Ergonomic heads with tilting and (optional) telescopic functions.

The improved CCIS Optical system includes a variety of contrast techniques like Fluorescence, dark field, polarization as well as an improved phase contrast: one condenser covers both positive as well as negative phase contrast lenses. While a solid-state quintuple nosepiece is standard, an optional sextuple nosepiece is now also available.

Imaging has also been improved through new CCD adapters, optimizing the use of all Motic Digital cameras with CMOS and CCD sensor targets.

The completely lead-free manufacturing of the microscope and its optics follow the RoHS regulations of environment and user protection.

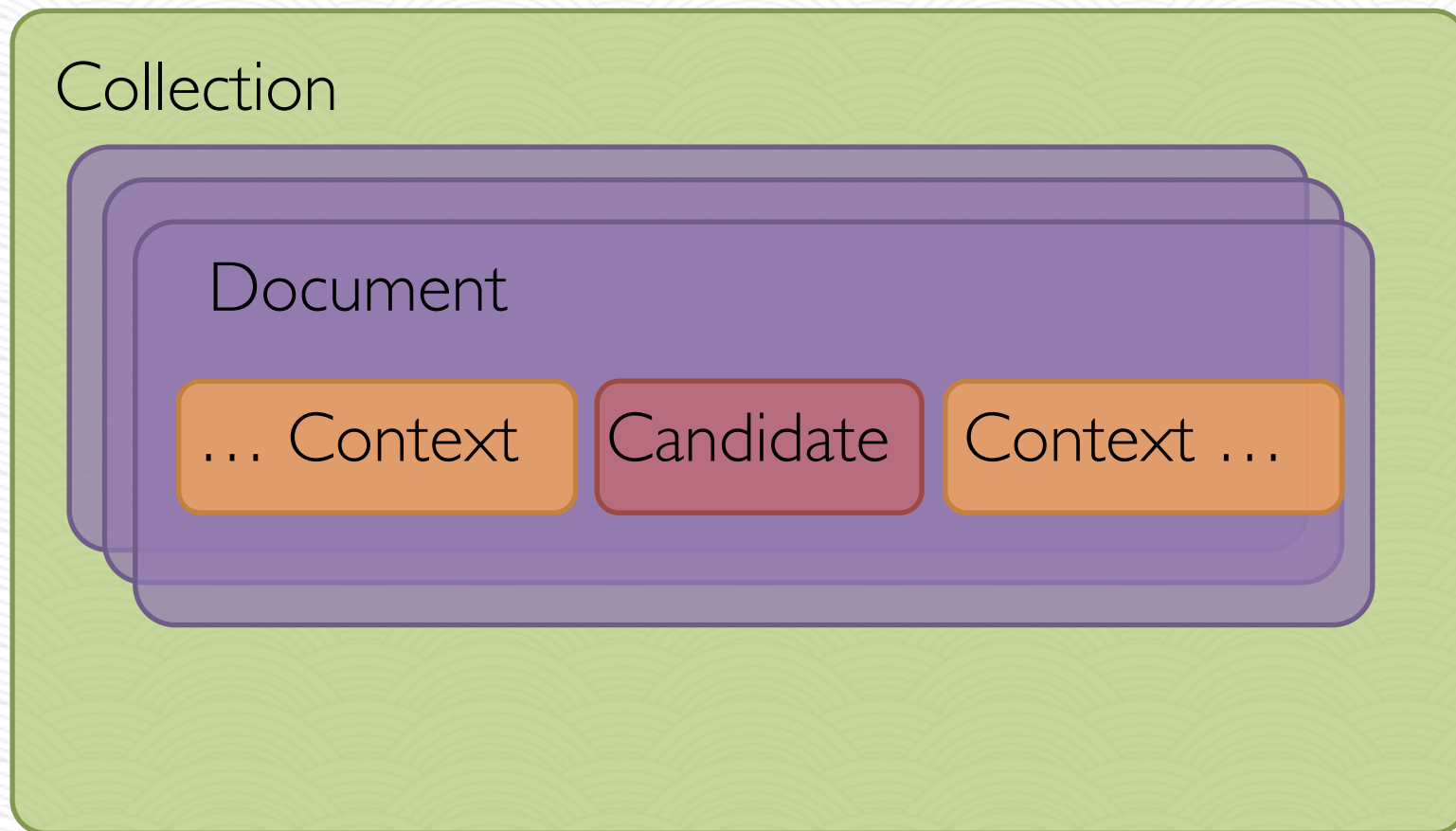
- 30° inclined Binocular head with 360° Swiveling eyepiece tubes for comfortable viewing while seated.
- Optionally binocular ergo head, tilting 4°~30° or binocular ergo plus head, tilting 4°~30° and telescoping 35mm
- Interpupillary distance adjustment between 48-75mm
- Widefield eyepieces N-WF10X/22mm with diopter adjustment on both eyepieces
- Reversed sextuple nosepiece with click stops for precise magnification changes
- CCIS EC-H Plan Achromatic objectives 4X/0.10, 10X/0.25, 40X/0.65 – Spring, 100X/1.25 – Spring/Oil
- Coaxial coarse and fine focusing system with 1 micron minimum increment with tension adjustment
- Vertical travel range 27mm
- Large 175mm X 145mm mechanical stage with low-position coaxial controls. Travel range 80 X 53mm. Sample holder can hold up to 2 slides
- Focusable and centrable Achromat swing-out condenser N.A. 0.90 with iris diaphragm
- Collector lens assembly with screw-on filter holder
- Koehler illumination quartz Halogen 6V/30W with external lamphouse and intensity control



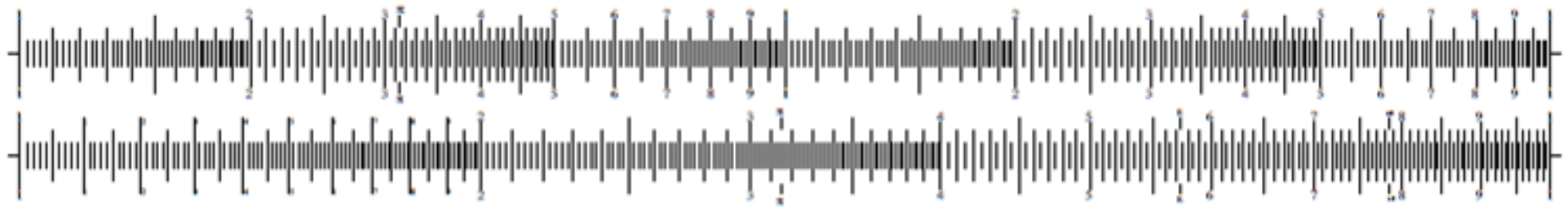
Courtesy: [Bing.com](https://www.bing.com)

Evidence has its scales

External



Sources for Keyphrase Evidence



SUMMARIZATION SCALES

Error Analysis

Directions Forward

Anthology as Platform

Scaling

NLP tackles “summarization” at different scales:

- Keywords
- Keyphrases
- Headlines
- Abstracts and summaries

Keywords / Phrases

- Position
 - Spread (Nguyen and Kan, 2007)
 - Section
- Structure
 - Part of Speech (Witten et al., 1999)
(Griveva, 2009)
(Hulth, 2004)
(Nakov, 2015)
 - N-gram Models (Liu et al., 2009)
(Nguyen and Phan, 2009)
(Wu et al., 2005)
- Supervision: Keyphraseness

Clauses – Headlines

HEADY (Alfonseca, 2013)

One-shot, single output

- Abstractive
- Pulling from multiple sites
- Density – Coverage
- Complexity – Penalty
 - Text simplification

Sentences - Summaries

- Predicate Structure
- Dependency Tuples
- Semantic Roles

- Redundancy
- Length penalty
- Cohesion

Summarization Facets

- Single vs. Multi
- Generic vs. Query-biased
- Stationary vs. Update
- Indicative vs. Informative

- Internal only vs. Leveraging External Resources

Table 1. *Participant summarization method features. tf: term frequency; loc: location; disc:discourse; coref: coreference; co-occ: co-occurrence; syn: synonyms.*

Participant	tf	loc	disc	coref	co-occ	syn
BT	+	+	-	+	+	-
CGI/CMU	+	+	-	-	+	-
CIR	+	+	-	-	-	+
Cornell/SabIR	+	-	-	-	+	-
GE	+	+	+	+	+	-
IA	+	-	-	-	+	-
IBM	+	+	-	-	-	-
ISI	+	+	-	-	-	+
LN	+	-	-	-	+	-
NMSU	+	-	+	+	-	-
NTU	+	-	+	+	-	-
Penn	-	+	-	+	-	-
SRA	+	+	-	+	-	+
Surrey	+	-	+	-	+	+
TextWise	+	-	-	+	+	+
UMass	+	-	-	-	+	-

Noise Reduction /
Signal Enhancement

cf (Erbs et al., 2015)

Mani et al. (2002)SUMMAC:
a text summarization
evaluation. Natural Language
Engineering 8(1). p 43-68.

Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering

Hongyuan Zha
Department of Computer
Science & Engineering
Pennsylvania State University
University Park, PA 16802
zha@cse.psu.edu

ABSTRACT

A novel method for *simultaneous* keyphrase extraction and generic text summarization is proposed by modeling text documents as weighted undirected and weighted bipartite graphs. Spectral graph clustering algorithms are used for partitioning sentences of the documents into topical groups with sentence link priors being exploited to enhance clustering quality. Within each topical group, saliency scores for keyphrases and sentences are generated based on a mutual reinforcement principle. The keyphrases and sentences are then ranked according to their saliency scores and selected for inclusion in the top keyphrase list and summaries of the document. The idea of building a hierarchy of summaries for documents capturing different levels of granularity is also briefly discussed. Our method is illustrated using several examples from news articles, news broadcast transcripts and web documents.

Categories and Subject Descriptors

1. INTRODUCTION

Text summarization is an increasingly pressing practical problem due to the explosion of the amount of textual information available. For example, web search engines have exploited the use of text summarization from the very beginning: starting with the extraction of certain number of bytes from the beginning of each document to the more sophisticated query-focused summaries typified by Google's snippets (see also the recent work in [1]). Query-focused summaries provide the users with the useful information for initial relevance judgement so that they can quickly zero in on documents deserving further inspection. In contrast, a generic summary in general distills the most important overall information from a document (or a set of documents), it can be especially useful when the documents are relatively long and contain a variety of topics. With many search engines starting to index documents in postscript and pdf formats, we will see increased availability of long and multi-part documents and the pressing needs for efficiently generating effective generic summaries for these documents. In addition,

Joint multi-resolution problem

- Math
- Nursing
- Software Engineering
- Literary Plays

Filter Settings for Current Results:

Read/Unread: All

Resource Type: All

Time Added: Since last login

Year of Publication: Any time

Sort by: Time Added

Profile:
[adult](#) [blood transfusion](#)
[cancer](#) [quality of life](#)
[pressure ulcer](#) [fall,time](#)
 ventilator-associated pneumonia
[\[cultural sampling\]](#)
 >>[Click to manage profile keywords](#)

Saved results:
 >>[Click to see saved results](#)

Hits 1-3 (out of about 3 total matching pages):

[Save marked results](#) [Mark as read](#) [Mark as unread](#)

[Impact of Invasive and Noninvasive Quantitative Culture Sampling on Outcome of Ventilator-Associated Pneumonia . A Pilot Study](#) -- SANCHEZ-NIETO et al. 157 (2): 371 -- American Journal of Respiratory and Critical Care Medicine
 Year of Publication: 1998, Full text, Added 2 days ago
 ... not given for pneumonia. In all cases ... specifically given for pneumonia. Twenty (83%) patients belonged to ... of late-onset (7 d) pneumonia. Late-onset pneumonia was considered in 14 ... directly attributable to pneumonia. This occurred in three ... mechanically ventilated patients with nosocomial pneumonia. In addition, quantitative ... mortality of ventilator-associated pneumonia (VAP) ranges from 20 to ... a poor outcome from nosocomial ...
<http://ajrccm.atsjournals.org/cgi/content/full/157/2/371> (cached) (key text)

[Prediction of Clinical Severity and Outcome of Ventilator-associated Pneumonia . Comparison of Simplified Acute Physiology Score with Systemic Inflammatory Mediators](#) -- FROON et al. 158 (4): 1026 -- American Journal of Respiratory and Critical Ca
 Year of Publication: 1998, Full text, Added 5 days ago
 ... Outcome of Ventilator-associated Pneumonia . Comparison of Simplified ... Outcome of Ventilator-associated Pneumonia Comparison of Simplified ... development of ventilator-associated pneumonia (VAP) (n = 42), diagnosed on ... RESULTS DISCUSSION REFERENCES Ventilator-associated pneumonia (VAP) is a frequently ... Definition of Ventilator-associated Pneumonia VAP was considered ICU-acquired ... the criteria for pneumonia developed after the patient ... clinical suspicion of pneumonia, bronchoscopy with bronchoalveolar lavage (BAL ...
<http://ajrccm.atsjournals.org/cgi/content/full/158/4/1026> (cached)

- Extracting Key Metadata

Intervention, Patient, Research Goal, Study Design

We performed an **open, prospective, randomized clinical trial** in 51 patients receiving mechanical ventilation for more than 72 h, in order to evaluate the impact of using **noninvasive (quantitative endotracheal aspirates [QEA])** diagnostic method on the morbidity and mortality of **ventilator-associated pneumonia (VAP)**.

Sex

Condition

Race

Age

Intervention

Study Design

Jin Zhao, Min-Yen Kan, Paula M. Procter, Siti Zubaidah, Wai Kin Yip and Goh Mien Li (2010)

[eEvidence: Information Seeking Support for Evidence-based Practice: An Implementation Case Study](#). In the Proceedings of the AMIA 2010 Annual Symposium. Washington, DC, USA.

Observations

Variations:

- Subjective, certainly no one annotator can capture all keywords, but perhaps would agree on the statehood
- However, we want good coverage: keyword set should covers all aspects of the item
- In some cases, the keyphrases are not part of formal metadata

Indicativeness

Def: “serving as a sign, indication or suggestion of something”

- Useful signpost of a category
- Discriminatory power (IDF)
- Represents the item to distinguish it from the corpus

Informativeness

Def: *“Providing information”*

- Importance within the document (TF)

Topicality

TF.IDF

Word in Context

- LDA
- Matrix Factorization
- Distributional approaches

Sources for Keyphrase Evidence
Summarization Scales



ERROR ANALYSIS

Directions Forward
Anthology as Platform

Hasan and Ng's (2014) error analysis

Canadian **Ben Johnson** left the **Olympics** today “in a complete state of shock,” accused of cheating with drugs in the world’s fastest **100-meter dash** and stripped of his **gold medal**. The prize went to American **Carl Lewis**. Many athletes accepted the accusation that Johnson used a muscle-building but dangerous and illegal anabolic steroid called **stanozolol** as confirmation of what they said they know has been going on in track and field. Two tests of Johnson’s urine sample proved positive and his denials of **drug use** were rejected today. “This is a blow for the **Olympic Games** and the **Olympic movement**,” said International Olympic Committee President Juan Antonio Samaranch.

- Overgeneration – Same keyword within different keyphrases
- Infrequency – Important but infrequently occurring
- Redundancy – Semantically equivalent output
- Evaluation – Evaluation metric problematic

Kazi S. Hasan and Vincent Ng. 2014.

[Automatic keyphrase extraction: A survey of the state of the art](#). In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1262–1273, Baltimore, Maryland, June. Association for Computational Linguistics

My take on Hasan and Ng (2014)

- Overgeneration – Same keyword within different keyphrases
 - Redundancy – Semantically equivalent output
 - Infrequency – Important but infrequently occurring
 - Evaluation – Evaluation metric problematic
- Cohort Effect* – consider candidates jointly
- Latent Category* – a priori knowledge informs keyphrase status

The Takeaways

Cohort Effect – consider candidates jointly

Latent Category – *a priori* knowledge informs keyphrase status

Abstractive Generalization – prefer a representative concept over concrete instances

Parts of a compound microscope



Taken from: moltic.com

Shop by category ▲

FREE SHIPPING on order

MICROSCOPE PACKAGES

Compound Microscopes

Stereo Microscopes

Digital Microscopes

Specialty Microscopes

Student Microscopes

Microscope Accessories

Magnifying Lamps

Microscope Cameras

Microscope Slides & Stains

Stereo Microscopes

Low power dissecting scopes

Clinical & Lab

Standard Lab/Clinical Stereo

CMO High Resolution

Home & Hobby

Kids

Hobbyist

Advanced

Industrial Inspection

Boom Stand Microscopes

Pedestal Microscopes

Platform Stands

Schools & Students

Elementary

Middle - High School

University

1117

Micr

AD

S /

ERA

ICE:

or

CA

Taken from: microscope.com

2.2 The Categories

To aid retrieval from database products such as Reuters Business Briefing (RBB), category codes from three sets (Topics, Industries, and Regions) were assigned to stories. The code sets were originally designed to meet customer requirements for access to corporate/business information, with the main focus on company coding and associated topics. With the introduction of the RBB product the focus broadened to the end user in large corporations, banks, financial services, consultancy, marketing, advertising and PR firms.

2.2.1 TOPIC CODES

Topic codes were assigned to capture the major subjects of a story. They were organized in four hierarchical groups: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). This code set provides a good example of how controlled vocabulary schemes represent a particular perspective on a data set. The RCV1 articles span a broad range of content, but the code set only emphasizes distinctions relevant to Reuters' customers. For instance, there are three different Topic codes for corporate ownership changes, but all of science and technology is a single category (GSCI).

2.2.2 INDUSTRY CODES

Industry codes were assigned based on types of businesses discussed in the story. They were grouped in 10 subhierarchies, such as I2 (METALS AND MINERALS) and I5 (CONSTRUCTION). The Industry codes make up the largest of the three code sets, supporting many fine distinctions.

1. Further formatting details are available at <http://about.reuters.com/researchandstandards/corpus/>.

```

5 <DOC ID="1">
6 <TITLE>2-Source Dispersers for Sub-Polynomial Entropy and Ramsey Graphs Beating the Frankl-Wilson Construction</TITLE>
7 <LINKS>
8 <LINK name="PDF" url="1/1.pdf"></LINK>
9 <LINK name="TXT" url="1/1.txt"></LINK>
10 <LINK name="HTML" url="1/1.html"></LINK>
11 <LINK name="XML" url="1/1.xml"></LINK>
12 </LINKS>
13 <CATEGORIES_AND_SUBJECT_DESCRIPTOR>
14 <ITEM>G.2.2 [Mathematics of Computing]: Discrete Mathematics - Graph algorithms</ITEM>
15 </CATEGORIES_AND_SUBJECT_DESCRIPTOR>
16 <GENERAL_TERMS>
17 <ITEM>Theory</ITEM>
18 <ITEM>Algorithms</ITEM>
19 </GENERAL_TERMS>
20 <AUTHOR_KEYWORDS>
21 <ITEM>Dispersers</ITEM>
22 <ITEM>Ramsey Graphs</ITEM>
23 <ITEM>Independent Sources</ITEM>
24 <ITEM>Extractors</ITEM>
25 </AUTHOR_KEYWORDS>
26 <KEYWORDS>
27 <KEYWORD_SET origin="0">
28 <ITEM>disperser</ITEM>
29 <ITEM>entropy</ITEM>
30 <ITEM>independent sources</ITEM>
31 <ITEM>extractor</ITEM>
32 <ITEM>randomness extraction</ITEM>
33 <ITEM>Ramsey graphs</ITEM>
34 <ITEM>bipartite graph</ITEM>
35 <ITEM>distribution</ITEM>
36 <ITEM>polynomial time computable disperser</ITEM>
37 <ITEM>subsource</ITEM>
38 <ITEM>subsource somewhere extractor</ITEM>
39 </KEYWORD_SET>
40 <KEYWORD_SET origin="1">
41 <ITEM>Ramsey graphs</ITEM>
42 <ITEM>extractors</ITEM>
43 <ITEM>disperser</ITEM>
44 <ITEM>construction of disperser</ITEM>
45 <ITEM>structure</ITEM>
46 <ITEM>tools</ITEM>
47 <ITEM>independent source</ITEM>
48 <ITEM>subsource</ITEM>
49 <ITEM>entropy</ITEM>
50 <ITEM>Theorem</ITEM>
51 <ITEM>resiliency</ITEM>
52 <ITEM>deficiency</ITEM>
53 </KEYWORD_SET>
54 <KEYWORD_SET origin="4">
55 <ITEM>explicit disperser</ITEM>
56 <ITEM>extractors</ITEM>
57 <ITEM>algorithms</ITEM>
58 <ITEM>Ramsey graph</ITEM>
59 <ITEM>sum-product theorem</ITEM>
60 <ITEM>block-sources</ITEM>
61 <ITEM>entropy</ITEM>
62 <ITEM>recursion</ITEM>
63 <ITEM>termination</ITEM>
64 <ITEM>resiliency</ITEM>
65 </KEYWORD_SET>

```

WINGNUS Keyphrase Corpus.

<http://wing.comp.nus.edu.sg/downloads/keyphraseCorpus/corpus.xml>

Sources for Keyphrase Evidence

Summarization Scales

Error Analysis



DIRECTIONS FORWARD

Anthology as Platform

Addressing the Takeaways

Cohort Effect – consider candidates jointly

- Redundancy / Entropy statistics
- Compound semantics

(Turney, 2003)
(Milhacea and Tarau, 2004)
(Boudin, 2015)

Latent Category – *a priori* knowledge

- Domain Named Entities
- Understanding the problem domain

Abstractive Generalization – extraction fails

- Exploit domain vocabularies
- Latent space (embeddings and exemplars)

Sparse data: an underlying problem

- Search space is very large
- Labeled observations are just alternatives

So make things more dense

- Project into a smaller space
- Select exemplars from their
- Consider their interactions

(Liu et al., 2015)

(Liu et al., 2009)

External Resources

- Scientific Documents
 - Citation Networks (Caragea et al., 2014)
(Gollapalli and Caragea, 2014)
 - Web Documents (Ferrara and Tasso, 2013)
 - Datastores (Freebase) (Marujo et al. 2013)
 - Wikipedia (Shi et al., 2008)
 - Query Log (Liang et al., 2009)
 - Social Media (Tuarob, 2015)
 - External Knowledgebase (Wu et al., 2005)

What's the purpose, anyways?

- For human vs. for machine process
- Inline highlight vs. Standalone
- Weighted (e.g., word cloud) vs. Presence
- Single vs. Multi (e.g., trend analysis)
- Generic vs. Query-biased (e.g., facets)

... and also language density.

But as for applications, we should be asking

What's the killer app for keyphrases?

WHAT IS BIG DATA?

VOLUME

Large amounts of data.

VELOCITY

Needs to be analyzed quickly.

VARIETY

Different types of structured and unstructured data.

Key questions enterprises are asking

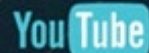
WHAT ARE THE VOLUMES OF DATA THAT WE ARE SEEING TODAY?



30 billion pieces of content were added to Facebook this past month by 600 million plus users.



Zynga processes 1 petabyte of content for players every day; a volume of data that is unmatched in the social game industry.



More than 2 billion videos were watched on YouTube... yesterday.



The average teenager sends 4,762 text messages per month.



32 billion searches were performed last month... on Twitter.

Source: Gartner

Everyday business and consumer life creates 2.5 quintillion bytes of data per day.

Trend Analytics for Social Media

WHAT DOES THE FUTURE LOOK LIKE?

Worldwide IP traffic will reach 1.5 zettabytes by 2015.



By 2015, 3 billion people

- Multimedia and social network evidence
- In the scholarly domain too

will be online, pushing the data created and shared to nearly 8 zettabytes.

HOW IS THE MARKET FOR BIG DATA SOLUTIONS EVOLVING?

A new IDC study says the market for big technology and services will grow from

Aman at Summer Palace Beijing

●●●●● 203 Reviews | #40 of 5,488 Hotels in Beijing | Certificate of Excellence

No.1 Gongmenqian Street, Yiheyuan, Haidian District | Summer Palace, Beijing 100091, China | Name/address in Chinese

Item Reviews

Enter dates for best prices

Check In Check Out

Show Prices



- Latent highlights
- Cohort effect
- Domain metadata and facets
- Online forums too



Kimcheemiso
Sydney
Level 3 Contributor
9 reviews
8 hotel reviews
15 helpful votes

“A piece of heaven money can buy”
●●●●● Reviewed June 22, 2015
It is a hotel like no other hotels, because it is actually a palace yet you don't feel it. The hotel compound is enormous, the rooms are big and decorated in the Imperial Chinese styles with all modern facilities and the restaurants serve good food. They have a staff/guest ratio of 4.5 to 1. Need..
More ▾
Was this review helpful?



“A nice experience, but I prefer Hangzhou more....”
●●●●● Reviewed June 22, 2015

Computational Advertising

Rocco Baldassarre

- Facebook: 74
- Twitter: 274
- Google+: 8
- LinkedIn: 1
- Pinterest: 58

407 SHARES
14904 READS

hold opposing views.

SEO And AdWords

SEO is a method that focuses on making your website content relevant to search engines. SEO or Search Engine Optimisation lend a hand for you to rank higher in the organic or natural search results. SEO optimizes your site in terms of keywords and articles at hand in your web pages. By optimizing these gears, search engines distinguish your site and put your site in the spots on the search results. SEO stands for Search Engine Optimisation. SEO is the process of getting more traffic to your website by getting the site listed and ranked highly for queries relating to your product, market or business in the natural or organic search results.

Adwords, a different internet marketing alternative, is advertising that is possible for you to position your advertisement on the top or right hand side of the search results pages on Google and other affiliate websites. You can target internet user's search queries and display your offers every time they look for it and you have budget left. The visitor then clicks on your advertisement and it takes them straightforwardly to your site. The advertiser will on the other hand pay each time somebody clicks on the advertisement, because AdWords runs on the pay per click system. With Adwords, your business listing can be displayed alongside the natural search results when people search for specific keyword phrases in Google. Search results displayed by Google Adwords are also called Pay Per Click (PPC) results.

- Keyphrases as friction
- Query logs, dialog as external evidence
- Query Expansion

Sources for Keyphrase Evidence

Summarization Scales

Error Analysis

Directions Forward



ANTHOLOGY AS PLATFORM

July 2015: This version of the ACL Anthology will become the default starting sometime this year. Click here to return to the previous version of the ACL Anthology. Both sites will be maintained in synchrony until the end of 2015. ✕

The Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing and its 15 associated workshops and events are now available in the Anthology. Also, the Proceedings of the Nineteenth Conference on Computational Natural Language Learning and its shared task, Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015), Proceedings of the 14th International Conference on Parsing Technologies (IWPT) and the Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015) are available on the ACL Anthology.

Welcome to the ACL Anthology

The ACL Anthology currently hosts 35531 papers on the study of computational linguistics and natural language processing. Subscribe to the mailing list to receive announcements and updates to the Anthology.

ACL Events	Present - 2010	2009 - 2000	1999 - 1990	1989 -
CL	15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84
TACL	15 14 13			
ACL	15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84
EACL	14 12	09 06 03	99 97 95 93 91	89 87 85
NAACL	15 13 12 10	09 07 06 04 03 01 00		
*SEMEVAL	15 14 13 12 10	07 04 01	98	
ANLP			97 94 92	88
EMNLP	14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96	
CONLL	15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97	
WS	15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 91 90	
SIGs	ANN BIOMED DAT DIAL FSM GEN HAN HUM LEX MEDIA MOL MT NLL PARSE MORPHON SEM SEM WAC			

Non-ACL Events	Present - 2010	2009 - 2000	1999 - 1990	1989 -
COLING	14 12 10	08 06 04 02 00	98 96 94 92 90	88
HLT	15 13 12 10	09 08 07 06 05 04 03 01	94 93 92 91 90	89
IJCNLP	15 13 11	09 08 05		



All Fields ▾

Search...

Search Q

Login

Bookmarks | History

Browse by

Author

Volume

Year

Venue

Attachment

Layers

MRF x

178

You searched for: Layers > MRF x

Start Over

« Previous | 1 - 10 of 178 | Next »

10 per page ▾

Sort by Relevance ▾

[P14-1007] Simple Negation Scope Resolution through Deep Parsing: A Semantic Solution to a Semantic Problem

Bookmark

📄 📄 🔍 Woodley Packard | Emily M. Bender | Jonathon Read | Stephan Oepen | Rebecca Dridan

[P14-1008] Logical Inference on Dependency-based Compositional Semantics

Bookmark

📄 📄 🔍 Ran Tian | Yusuke Miyao | Takuya Matsuzaki

[P14-1009] A practical and linguistically-motivated approach to compositional distributional semantics

Bookmark

📄 📄 🔍 Denis Paperno | Nghia The Pham | Marco Baroni

[P14-1010] Lattice Desegmentation for Statistical Machine Translation

Bookmark

📄 📄 🔍 Mohammad Salameh | Colin Cherry | Grzegorz Kondrak

[P14-1011] Bilingually-constrained Phrase Embeddings for Machine Translation

Bookmark

📄 📄 🔍 Jiajun Zhang | Shujie Liu | Mu Li | Ming Zhou | Chenqinq Zong



All Fields ▾

Search...

Search Q

Login
Bookmarks | History

Anthology: P14-1007
Volume: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
Authors: Woodley Packard | Emily M. Bender | Jonathon Read | Stephan Oepen | Rebecca Dridan
Month: June
Year: 2014
Venue: ACL
Address: Baltimore, Maryland
SIG:
Publisher: Association for Computational Linguistics
Pages: 69-78
URL: <http://aclweb.org/anthology/P14-1007>
DOI: [10.3115/v1/P14-1007](https://doi.org/10.3115/v1/P14-1007)
MRF: LaTeXXML
Bibtype: inproceedings
Bibkey: packard-EtAl:2014:P14-1

Bib Export formats: [BibTeX](#) [RIS](#) [Endnote](#) [MODS XML](#) [MS Word '07](#)

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```

<!--
XML document generated using OCR technology from Nuance Communications, Inc.
-->
<document xmlns="http://www.scansoft.com/omnipage/xml/ssdoc-schema3.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <page ocr-vers="OmniPageCSDK18" app-vers="OmniPageCSDK18">
    <description>
      <source file="C://Users...>
      <theoreticalPage size=...>
      <width="11918" height="...>
      <language>en</language>
    </description>
    <body>
      <section l="936" t="118...>
        <column l="936" t="11...>
          <para l="2467" t="1...>
            <ln l="2467" t="1...>
              <column l="936" t="1180" r="11016" b="3808">
                <para l="2467" t="1253" r="9470" b="1781" alignment="centered">
                  <ln l="2467" t="1253" r="9470" b="1517" baseLine="1450" bold=
                    fontFace="Times New Roman" fontFamily="roman" fontPitch="vari
                      <wd l="2467" t="1253" r="3298" b="1517">Simple</wd>
                      <space/>
                      <wd l="3374" t="1253" r="4474" b="1517">Negation</wd>
                      <space/>
                      <wd l="4555" t="1253" r="5261" b="1517">Scope</wd>
                      <space/>
                      <wd l="5338" t="1253" r="6638" b="1459">Resolution</wd>
                      <space/>
                      <wd l="6720" t="1258" r="7690" b="1517">through</wd>
                      <space/>
                      <wd l="7766" t="1258" r="8386" b="1517">Deep</wd>
                      <space/>
                    </ln>
                  <ln l="3302" t="1570" r="8645" b="1781" baseLine="1771" bold="true" underlined="none" superscript="none" fontSize="1450"
                    fontFace="Times New Roman" fontFamily="roman" fontPitch="variable" spacing="0" forcedEOF="true">
                      <wd l="3302" t="1570" r="3504" b="1771">A</wd>
                      <space/>
                      <wd l="3590" t="1570" r="4709" b="1781">Semantic</wd>
                      <space/>
                      <wd l="4790" t="1570" r="5803" b="1781">Solution</wd>
                      <space/>
                      <wd l="5880" t="1589" r="6110" b="1776">to</wd>
                      <space/>
                      <wd l="6192" t="1632" r="6331" b="1776">a</wd>
                      <space/>
                      <wd l="6408" t="1570" r="7531" b="1781">Semantic</wd>
                      <space/>
                      <wd l="7603" t="1574" r="8645" b="1776">Problem</wd>
                    </ln>
                  </para>
                </column>
              </ln>
            </section>
          </column>
        </section>
      </body>
    </page>
  </document>

```

OmniPage Commercial OCR

- Spatial location and font properties
- Reading order resolved

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<algorithms version="110505">
  ▼<algorithm name="SectLabel" version="110505">
    ▼<variant no="0" confidence="0.000017">
      ▼<title confidence="0.997781">
        Simple Negation Scope Resolution through Deep Parsing: A Semantic Solution to a Semantic Problem
      </title>
      ▼<author confidence="0.999325">
        Woodley Packard4, Emily M. Bender4, Jonathon Read*, Stephan Oepen°Q, and Rebecca Dridan°
      </author>
      ▼<affiliation confidence="0.98256975">
        ♦ University of Washington, Department of Linguistics ♦ Teesside University, School of Computing ▼ University of Oslo, Department of Informatics ♦ Potsdam University, Department of Linguistics
      </affiliation>
      ▼<email confidence="0.984932">
        ebender@uw.edu, sweaglesw@sweaglesw.org, j.read@tees.ac.uk, { oe |rdridan }@ifi.uio.no
      </email>
```

results on this task to date.

</bodyText>

<sectionHeader confidence="0.998991" genericHeader="introduction">1 Introduction</sectionHeader>

▼<bodyText confidence="0.999919785714286">

Recently, there has been increased community interest in the theoretical and practical analysis of what Morante and Sporleder (2012) call modality and negation, i.e. linguistic expressions that modulate the certainty or factuality of propositions. Automated analysis of such aspects of meaning is important for natural language processing tasks which need to consider the truth value of statements, such as for example text mining (Vincze et al., 2008) or sentiment analysis (Lapponi et al., 2012). Owing to its immediate utility in the curation of scholarly results, the analysis of negation and so-called hedges in bio-medical research literature has been the focus of several workshops, as well as the Shared Task at the 2011 Conference on Computational Linguistics (CoNLL). Task 1 at the First Joint Conference on Lexical and Computational Semantics (EMNLP) (2012) provided a fresh, principled annotation of negation and called for systems to analyze negation-detecting cues (affixes, words, or phrases that express negation), resolving their scopes (which parts of the sentence they apply to) and identifying the negated event or property. The task organizers designed and applied it to a little more than 100,000 tokens of running text by the novelist Sir Arthur Conan Doyle. While the task participating systems were evaluated on a purely compositional semantics (Basile et al., 2012), with results ranking in the middle of the 12 participating systems. Conversely, the best-performing systems approached the task through machine learning or heuristic processes over relatively coarse-grained representations; see § 2 below. Example (1), where O marks the cue and {} the in-scope elements, illustrates the annotations, including how negation inside a noun phrase can scope over discontinuous parts of the sentence.1

</bodyText>

<sectionHeader confidence="0.998991" genericHeader="introduction">1 Introduction</sectionHeader>

▼<bodyText confidence="0.999919785714286">

Recently, there has been increased community interest in the theoretical and practical analysis of what Morante and Sporleder (2012) call modality and negation, i.e. linguistic expressions that modulate the certainty or factuality of propositions. Automated analysis of such aspects of meaning is important for natural language processing tasks which need to consider the truth value of statements, such as for example text mining (Vincze et al., 2008) or sentiment analysis (Lapponi et al., 2012). Owing to its immediate utility in the curation of scholarly results, the analysis of negation and so-called hedges in bio-medical research literature has been the focus of several workshops, as well as the Shared Task at the 2011 Conference on Computational Linguistics (CoNLL). Task 1 at the First Joint Conference on Lexical and Computational Semantics (EMNLP) (2012) provided a fresh, principled annotation of negation and called for systems to analyze negation-detecting cues (affixes, words, or phrases that express negation), resolving their scopes (which parts of the sentence they apply to) and identifying the negated event or property. The task organizers designed and applied it to a little more than 100,000 tokens of running text by the novelist Sir Arthur Conan Doyle. While the task participating systems were evaluated on a purely compositional semantics (Basile et al., 2012), with results ranking in the middle of the 12 participating systems. Conversely, the best-performing systems approached the task through machine learning or heuristic processes over relatively coarse-grained representations; see § 2 below. Example (1), where O marks the cue and {} the in-scope elements, illustrates the annotations, including how negation inside a noun phrase can scope over discontinuous parts of the sentence.1

</bodyText>

ParsCit Document Segmentation

- Retrieves body text and categorizes other text

- Resolves headers to

generic header category

```

▼<citation valid="true">
  ▼<authors>
    <author>H Alshawi</author>
  </authors>
  <title>The Core Language Engine.</title>
  <date>1992</date>
  <publisher>MIT Press.</publisher>
  <location>Cambridge, MA, USA:</location>
  ▼<contexts>
    ▼<context position="12000" citStr="Alshawi, 1992" startWordPosition="1857" endWordPosition="1858">
      k annotations. 3.1 MRS Crawling Fig. 1 shows the ERG semantic analysis for our running example. The heart of the MRS is a
      multiset of elementary predications (EPs). Each ele4Read et al. (2012) predicted cues using a closed vocabulary assumption with a
      supervised classifier to disambiguate instances of cues. 5In other words, a possible semantic interpretation of the (string-
      based) Shared Task annotation guidelines and data is in terms of a quantifier-free approach to meaning representation, or in
      others, Alshawi, 1992). From this
      compass interactions of negation
      'andle', prefixed to predicates with a
      Eventualities (ei) in MRS denote
      or abstract) entities. All EPs have the
    </context>
  </contexts>
  ▼<citation valid="true">
    ▼<authors>
      <author>H Alshawi</author>
    </authors>
    <title>The Core Language Engine.</title>
    <date>1992</date>
    <publisher>MIT Press.</publisher>
    <location>Cambridge, MA, USA:</location>
    ▼<contexts>
      ▼<context position="12000" citStr="Alshawi, 1992" startWo
        k annotations. 3.1 MRS Crawling Fig. 1 shows the ERG s
        multiset of elementary predications (EPs). Each ele4Re
        supervised classifier to disambiguate instances of cue
        based) Shared Task annotation guidelines and data is i
        terms of one where quantifier scope need not be made e
        interpretation, it follows that the notion of scope as
      </context>
    </contexts>
  </citation>
  </title>
  <date>2012</date>
  ▼<booktitle>
    In Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (p. 301–309).
  </booktitle>
  <location>Montréal, Canada.</location>
  ▼<contexts>
    ▼<context position="2978" citStr="Basile et al., 2012" startWordPosition="445" endWordPosition="525">
      of negation and called for systems to analyze negation-detecting cues (affixes, words, or phrases that express negation),
      resolving their scopes (which parts of a sentence are actually negated), and identifying the scope of the negation. The task
      organizers designed and documented an annotation scheme (Morante and Daelmans, 2012) and applied it to a little more than
      100,000 tokens of running text by the novelist Sir Arthur Conan Doyle. While the task was framed from a semantic
      perspective, only one participating system actually employed explicit compositional semantics (Basile et al., 2012), with results
      ranking in the middle of the 12 participating systems. Conversely, the best-performing systems approached the task through machine
      learning or heuristic processing over syntactic and linguistically relatively coarse-grained representations; see § 2 below.
      Example (1), where 0 marks the cue and /1 the in-scope elements, illustrates the annotations, including how negation inside a
    </context>
  </contexts>

```

ParsCit Citation Parsing
 - CRF based reference string
 with auto detected citation
 context (citance)

Anthology: P14-1007
Volume: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
Authors: Woodley Packard | Emily M. Bender | Jonathon Read | Stephan Oepen | Rebecca Dridan
Month: June
Year: 2014
Venue: ACL
Address: Baltimore, Maryland
SIG:
Publisher: Association for Computational Linguistics
Pages: 69-78
URL: <http://aclweb.org/anthology/P14-1007>
DOI: 10.3115/v1/P14-1007
MRF: LaTeXML
Omni-OCR
ParsCit-Text
XX-Keyphrase
YY- Summary

Bibtype: inproceedings
Bibkey: packard-EtAl:2014:P14-1
Bib Export formats: [BibTeX](#) [RIS](#) [Endnote](#) [MODS XML](#) [MS Word '07](#)

Featuring your work in the future?

Long term: Needs an API-based method for accepting new text for processing

2016: Shared task on the Anthology?


Previous:

- CL Pilot Summarization Task at TAC 2014:
(Jaidka et al., 14) <https://github.com/WINGNUS/scisumm-corpus>

Now planning:

- Which tasks are of interest to the community?
 - Keyphrase
 - Summarization
- What venue is the best opportunity?
 - An ACL workshop?
- What role could you commit to participate as?

Conclusion

- Larger summarization scales can inform our task
- Errors stemming from a cohort effect, latent categories and abstractive generalizations
- Characteristics of the keyphrase application may also inform
- Call for Participation: 
For scholarly text, let's start with our own text

Thank you!