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
SOURCES FOR KEYPHRASE EVIDENCE

Image from microscope.com

Slides available at: dwz.cn/kan-kp
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
Evidence has its scales

External

Collection

Document

… Context  Candidate  Context …
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
  – Section

• Structure
  – Part of Speech
  – N-gram Models

• Supervision: Keyphraseness

(Nguyen and Kan, 2007)
(Witten et al., 1999)
(Griveva, 2009)
(Hulth, 2004)
(Nakov, 2015)
(Liu et al., 2009)
(Nguyen and Phan, 2009)
(Wu et al., 2005)
Clauses – Headlines

*One-shot, single output*

- Abstractive
- Pulling from multiple sites

- Density – Coverage
- Complexity – Penalty
  - Text simplification

HEADY (Alfonseca, 2013)
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.

<table>
<thead>
<tr>
<th>Participant</th>
<th>tf</th>
<th>loc</th>
<th>disc</th>
<th>coref</th>
<th>co-occ</th>
<th>syn</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>CGI/CMU</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>CIR</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Cornell/SabIR</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>GE</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>IA</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>IBM</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ISI</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>LN</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>NMSU</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NTU</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Penn</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SRA</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Surrey</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>TextWise</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>UMass</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

Noise Reduction / Signal Enhancement

cf (Erbs et al., 2015)

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.

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 those documents. In addition,
Joint multi-resolution problem

- Math
- Nursing
- Software Engineering
- Literary Plays

Extracting Key Metadata

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).

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
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
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

- eyepiece
- objective lens
- stage
- illuminator
- base
- focusing knob
- revolving nosepiece
- stage clip
- disc diaphragm
- arm
- Inclined head

Taken from: moltic.com
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

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/.

27 Jul 2015

Keyphrase Workshop

WINGNUS Keyphrase Corpus.
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

*Latent Category* – *a priori* knowledge
- Domain Named Entities
- Understanding the problem domain

*Abstractive Generalization* – extraction fails
- Exploit domain vocabularies
- Latent space (embeddings and exemplars)

(Turney, 2003)
(Milhacea and Tarau, 2004)
(Boudin, 2015)
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
  - Web Documents
  - Datastores (Freebase)
  - Wikipedia
  - Query Log
  - Social Media
  - External Knowledgebase

(Caragea et al., 2014)
(Gollapalli and Caragea, 2014)
(Ferrara and Tasso, 2013)
(Marujo et al. 2013)
(Shi et al., 2008)
(Liang et al., 2009)
(Tuarob, 2015)
(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?
Trend Analytics for Social Media

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

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

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

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

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 terms of keywords and articles at hand in your web pages. By these gears, search engines distinguish your site and put your spots on the search results. SEO stands for Search Engine Optimisation. SEO is the process of getting more traffic to your website by listed and ranked highly for queries relating to your product, market business in the natural or organic search results.

Adwords, a different internet marketing alternative, is advertising possible for you to position your advertisement on the top or target internet user’s search queries and display your offers. You 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.

Courtesy: searchengineland.com
Sources for Keyphrase Evidence
Summarization Scales
Error Analysis
Directions Forward

ANTHOLOGY AS PLATFORM
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.

**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.
You searched for: Layers > MRF

1 - 10 of 178

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

[P14-1008] Logical Inference on Dependency-based Compositional Semantics
- Ran Tian | Yusuke Miyao | Takuya Matsuzaki

[P14-1009] A practical and linguistically-motivated approach to compositional distributional semantics
- Denis Paperno | Nghia The Pham | Marco Baroni

[P14-1010] Lattice Desegmentation for Statistical Machine Translation
- Mohammad Salameh | Colin Cherry | Grzegorz Kondrak

- Jiatun Zhang | Shutie Liu | Mu Li | Ming Zhou | Chenqing Zong
OmniPage Commercial OCR
- Spatial location and font properties
- Reading order resolved
Recently, there has been increased 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 (Lapata et al., 2011). Owing to its immediate utility in the cura-tion of scholarly results, the analysis liter-ature has been the focus of several workshops, as well as the Shared Learning (CoML). Task 1 at the First Joint Conference on Lex-ical and CoEx-2012 provided a fresh, pruned-gaged annotation of negation and called for the tagging of negation and so-called hedge/statement parts of text. Off the text mining and noma-ics, the task setting was intended to analyze decision-making rules (affirmation, negation, and modality) in scholarly papers. Participants were asked to identify the negated event or property. The task organizers designed a task and applied it to a little more than 100,000 tokens of running text and annotations were framed from a semantic perspective only one particle semantics (Basile et al., 2012), with results ranked in the middle of the performing systems. The task involved the task through machine learning or heuristic relative coarse-grained representations; see § 2 below. Example (1), which illustrates the annotations, including how negation inside a noun phrase co-

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

CRF based reference string with auto detected citation context (citance)

Example (1), where $O$ marks the cue and $\{\}$ the in-scope elements, illustrates the annotations, including how negation inside a
Featuring your work in the future?

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!