SemEval-2010 Task 5:

Automatic Keyphrase Extraction from Scientific Articles

Su Nam Kim[1], Olena Medelyan[2], Min-Yen Kan[3] and Timothy Baldwin[1]

[1] The University of Melbourne
[2] Pingar LP
[2] National University of Singapore

15 July 2010

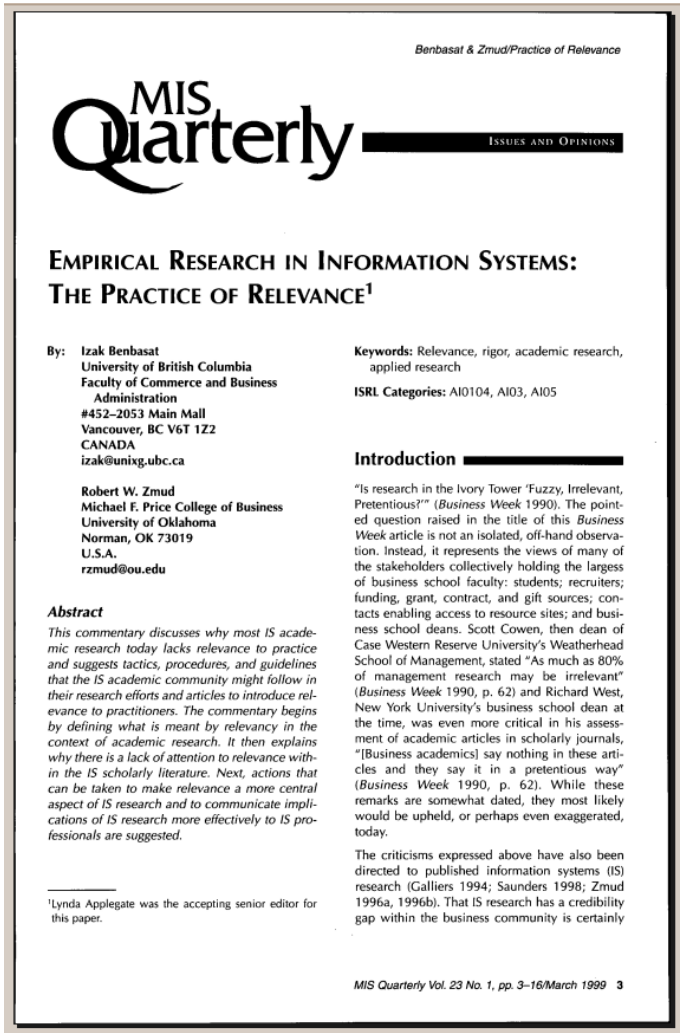
Overview

Keyphrases represent the main topics in articles

Our Goal:

- Offer systems an opportunity to compete comparably:
  - rank systems and approaches;
  - ascertain successful techniques;
  - investigate effectiveness on different subdomains.
- Generate a standard data set for future research.

Overview



Benbasat & Zmud/Practice of Relevance

MIS Quarterly — Issues and Opinions

Empirical Research in Information Systems: The Practice of Relevance[1]

By: Izak Benbasat
University of British Columbia
Faculty of Commerce and Business Administration
#452–2053 Main Mall
Vancouver, BC V6T 1Z2
CANADA
izak@unixg.ubc.ca

Robert W. Zmud
Michael F. Price College of Business
University of Oklahoma
Norman, OK 73019
U.S.A.
rzmud@ou.edu

Abstract

This commentary discusses why most IS academic research today lacks relevance to practice and suggests tactics, procedures, and guidelines that the IS academic community might follow in their research efforts and articles to introduce relevance to practitioners. The commentary begins by defining what is meant by relevancy in the context of academic research. It then explains why there is a lack of attention to relevance within the IS scholarly literature. Next, actions that can be taken to make relevance a more central aspect of IS research and to communicate implications of IS research more effectively to IS professionals are suggested.

[1]Lynda Applegate was the accepting senior editor for this paper.

Keywords: Relevance, rigor, academic research, applied research

ISRL Categories: AI0104, AI03, AI05

Introduction

"Is research in the Ivory Tower 'Fuzzy, Irrelevant, Pretentious?'" (*Business Week* 1990). The pointed question raised in the title of this *Business Week* article is not an isolated, off-hand observation. Instead, it represents the views of many of the stakeholders collectively holding the largess of business school faculty: students; recruiters; funding, grant, contract, and gift sources; contacts enabling access to resource sites; and business school deans. Scott Cowen, then dean of Case Western Reserve University's Weatherhead School of Management, stated "As much as 80% of management research may be irrelevant" (*Business Week* 1990, p. 62) and Richard West, New York University's business school dean at the time, was even more critical in his assessment of academic articles in scholarly journals, "[Business academics] say nothing in these articles and they say it in a pretentious way" (*Business Week* 1990, p. 62). While these remarks are somewhat dated, they most likely would be upheld, or perhaps even exaggerated, today.

The criticisms expressed above have also been directed to published information systems (IS) research (Galliers 1994; Saunders 1998; Zmud 1996a, 1996b). That IS research has a credibility gap within the business community is certainly

$\Rightarrow$ *relevance, rigor, academic research, applied research*

Overview (2)

Difficulties in Automatic Keyphrase Extraction

- Identification of valid terms (*candidate selection*; i.e., NN, NP);
- Dealing with lexical variation (*candidate comparison/paraphrasing*);
- specification vs. generalization (*ranking candidates*).

Notion of Significance used in many NLP applications

- Semantic metadata for summarization (Barzilay 1997, Lawrie 2001, DAvanzo 2005)
- Document indexing (Gutwin 1999)
- Document clustering (Zhang 2004, Hammouda 2005)
- Document summarization (Berger 2000, Buyukkokten 2001)

Existing Keyphrase Corpora

We note there are already some publicly-available data sets (inter alia):

- 2,000 journal abstracts from Inspec (Hulth 2004)
- 120 documents from ACM Library (Nguyen 2007)
- 308 documents from DUC 2001 (Wan 2008)
- 1,323 documents from PubMed (Schutz 2008)
- 180 documents from CiteULike.org, multiple sets per doc (Medelyan 2009)

The SemEval Task 5 Dataset

We specifically target scholarly computer science articles.

- 284 conference & workshop papers from the ACM Digital Library
- 4 1998 ACM classification areas, purposefully different:
  - C2.4 *Distributed Systems*
  - H3.3 *Information Search & Retrieval*
  - I2.11 *Distributed Artificial Intelligence – Multiagent Systems*
  - J4 *Social and Behavioral Sciences – Economics*
- 6–7 pages, including tables & figures
- 40 trial, 144 training and 100 test documents

Document Distribution

Strove for uniform distribution w.r.t. categories and dataset splits:

Dataset	Total	Document Area			
		Distr. Systems	IR	AI	Social Sci.
Trial	40	10	10	10	10
Training	144	34	39	35	36
Test	100	25	25	25	25

Table: Number of documents per ACM classifications area in each dataset

Annotation to the Corpus

- 50 volunteer students from the CS department of NUS
  (unaffiliated with the NUS participation effort team)
- 5 papers per annotator, up to 15 keyphrases per paper
- Accepted variations:
  1. *A of B* $\rightarrow$ *B A* (e.g. *policy of school* $=$ *school policy*)
  2. *A's B* $\rightarrow$ *A B* (e.g. *school's policy* $=$ *school policy*)
     cf. some exceptions (e.g. *matter of fact* vs. *?fact matter*).
- Averages and Salient Statistics
  - 4 author- and 12 reader-assigned keyphrases per doc
  - 77.8% author-assigned keyphrases matched reader-assigned ones
  - 19% author- and 15% reader-assigned keyphrases not found in text

Keyphrase Distribution

Again, we strove for uniform distribution:

Dataset	Author	Reader	Combined
Trial	150	500	600
Training	560	1800	2200
Test	390	1200	1500

Table: Approximate number of author- and reader-assigned keyphrases in each dataset split

Evaluation Metrics and Baselines

- Metric: Micro-averaged precision, recall & F-score by exact matching at top 5, 10 and 15 ranks
- Baselines:
  - Unsupervised: top *n*–grams ranked by TF·IDF
  - Supervised: Naïve Bayes (NB) & Maximum Entropy (ME) classifiers, TF·IDF-weighted term features

Method	Source	Top 5 candidates			Top 10 candidates			Top 15 candidates		
		P	R	F	P	R	F	P	R	F
TF·IDF	Reader	17.8%	7.4%	10.4%	13.9%	11.5%	12.6%	11.6%	14.5%	12.9%
	Combined	22.0%	7.5%	11.2%	17.7%	12.1%	14.4%	14.9%	15.3%	15.1%
NB	Reader	16.8%	7.0%	9.9%	13.3%	11.1%	12.1%	11.4%	14.2%	12.7%
	Combined	21.4%	7.3%	10.9%	17.3%	11.8%	14.0%	14.5%	14.9%	14.7%
ME	Reader	16.8%	7.0%	9.9%	13.3%	11.1%	12.1%	11.4%	14.2%	12.7%
	Combined	21.4%	7.3%	10.9%	17.3%	11.8%	14.0%	14.5%	14.9%	14.7%

Table: Baseline keyphrase extraction performance

Performance on combined keyphrases

System		Top 5 candidates			Top 10 candidates			Top 15 candidates		
		P	R	F	P	R	F	P	R	F
HUMB	1	39.0%	13.3%	19.8%	32.0%	21.8%	26.0%	27.2%	27.8%	**27.5%**
WINGNUS	2	40.2%	13.7%	20.5%	30.5%	20.8%	24.7%	24.9%	25.5%	25.2%
KP-Miner	3	36.0%	12.3%	18.3%	28.6%	19.5%	23.2%	24.9%	25.5%	25.2%
SZTERGAK	4	34.2%	11.7%	17.4%	28.5%	19.4%	23.1%	24.8%	25.4%	25.1%
ICL	5	34.4%	11.7%	17.5%	29.2%	19.9%	23.7%	24.6%	25.2%	24.9%
SEERLAB	6	39.0%	13.3%	19.8%	29.7%	20.3%	24.1%	24.1%	24.6%	24.3%
KX_FBK	7	34.2%	11.7%	17.4%	27.0%	18.4%	21.9%	23.6%	24.2%	23.9%
DERIUNLP	8	27.4%	9.4%	13.9%	23.0%	15.7%	18.7%	22.0%	22.5%	22.3%
MAUI	9	35.0%	11.9%	17.8%	25.2%	17.2%	20.4%	20.3%	20.8%	20.6%
DFKI	10	29.2%	10.0%	14.9%	23.3%	15.9%	18.9%	20.3%	20.7%	20.5%
BUAP	11	13.6%	4.6%	6.9%	17.6%	12.0%	14.3%	19.0%	19.4%	19.2%
SJTULTLAB	12	30.2%	10.3%	15.4%	22.7%	15.5%	18.4%	18.4%	18.8%	18.6%
UNICE	13	27.4%	9.4%	13.9%	22.4%	15.3%	18.2%	18.3%	18.8%	18.5%
UNPMC	14	18.0%	6.1%	9.2%	19.0%	13.0%	15.4%	18.1%	18.6%	18.3%
JU_CSE	15	28.4%	9.7%	14.5%	21.5%	14.7%	17.4%	17.8%	18.2%	18.0%
LIKEY	16	29.2%	10.0%	14.9%	21.1%	14.4%	17.1%	16.3%	16.7%	16.5%
UvT	17	24.8%	8.5%	12.6%	18.6%	12.7%	15.1%	14.6%	14.9%	14.8%
POLYU	18	15.6%	5.3%	7.9%	14.6%	10.0%	11.8%	13.9%	14.2%	14.0%
UKP	19	9.4%	3.2%	4.8%	5.9%	4.0%	4.8%	5.3%	5.4%	5.3%

Table: Ranked by F_1@15

Performance on reader keyphrases

System		Top 5 candidates			Top 10 candidates			Top 15 candidates		
		P	R	F	P	R	F	P	R	F
HUMB	1	30.4%	12.6%	17.8%	24.8%	20.6%	22.5%	21.2%	26.4%	**23.5%**
KX_FBK	2	29.2%	12.1%	17.1%	23.2%	19.3%	21.1%	20.3%	25.3%	22.6%
SZTERGAK	3	28.2%	11.7%	16.6%	23.2%	19.3%	21.1%	19.9%	24.8%	22.1%
WINGNUS	4	30.6%	12.7%	18.0%	23.6%	19.6%	21.4%	19.8%	24.7	22.0%
ICL	5	27.2%	11.3%	16.0%	22.4%	18.6%	20.3%	19.5%	24.3%	21.6%
SEERLAB	6	31.0%	12.9%	18.2%	24.1%	20.0%	21.9%	19.3%	24.1%	21.5%
KP-Miner	7	28.2%	11.7%	16.5%	22.0%	18.3%	20.0%	19.3%	24.1%	21.5%
DERIUNLP	8	22.2%	9.2%	13.0%	18.9%	15.7%	17.2%	17.5%	21.8%	19.5%
DFKI	9	24.4%	10.1%	14.3%	19.8%	16.5%	18.0%	17.4%	21.7%	19.3%
UNICE	10	25.0%	10.4%	14.7%	20.1%	16.7%	18.2%	16.0%	19.9%	17.8%
SJTULTLAB	11	26.6%	11.1%	15.6%	19.4%	16.1%	17.6%	15.6%	19.4%	17.3%
BUAP	12	10.4%	4.3%	6.1%	13.9%	11.5%	12.6%	14.9%	18.6%	16.6%
MAUI	13	25.0%	10.4%	14.7%	18.1%	15.0%	16.4%	14.9%	18.5%	16.1%
UNPMC	14	13.8%	5.7%	8.1%	15.1%	12.5%	13.7%	14.5%	18.0%	16.1%
JU_CSE	15	23.4%	9.7%	13.7%	18.1%	15.0%	16.4%	14.4%	17.9%	16.0%
LIKEY	16	24.6%	10.2%	14.4%	17.9%	14.9%	16.2%	13.8%	17.2%	15.3%
POLYU	17	13.6%	5.7%	8.0%	12.6%	10.5%	11.4%	12.0%	14.9%	13.3%
UvT	18	20.4%	8.5%	12.0%	15.6%	13.0%	14.2%	11.9%	14.9%	13.2%
UKP	19	8.2%	3.4%	4.8%	5.3%	4.4%	4.8%	4.7%	5.8%	5.2%

Table: Ranked by F_1@15

Performance on author keyphrases

System	R	Top 5 candidates			Top 10 candidates			Top 15 candidates		
		P	R	F	P	R	F	P	R	F
HUMB	1	21.2%	27.4%	23.9%	15.4%	39.8%	22.2%	12.1%	47.0%	**19.3%**
KP-Miner	2	19.0%	24.6%	21.4%	13.4%	34.6%	19.3%	10.7%	41.6%	17.1%
ICL	3	17.0%	22.0%	19.2%	13.5%	34.9%	19.5%	10.5%	40.6%	16.6%
MAUI	4	20.4%	26.4%	23.0%	13.7%	35.4%	19.8%	10.2%	39.5%	16.2%
SEERLAB	5	18.8%	24.3%	21.2%	13.1%	33.9%	18.9%	10.1%	39.0%	16.0%
SZTERGAK	6	14.6%	18.9%	16.5%	12.2%	31.5%	17.6%	9.9%	38.5%	15.8%
WINGNUS	7	18.6%	24.0%	21.0%	12.6%	32.6%	18.2%	9.3%	36.2%	14.8%
DERIUNLP	8	12.6%	16.3%	14.2%	9.7%	25.1%	14.0%	9.3%	35.9%	14.7%
KX_FBK	9	13.6%	17.6%	15.3%	10.0%	25.8%	14.4%	8.5%	32.8%	13.5%
BUAP	10	5.6%	7.2%	6.3%	8.1%	20.9%	11.7%	8.3%	32.0%	13.2%
JU_CSE	11	12.0%	15.5%	13.5%	8.5%	22.0%	12.3%	7.5%	29.0%	11.9%
UNPMC	12	7.0%	9.0%	7.9%	7.7%	19.9%	11.1%	7.1%	27.4%	11.2%
DFKI	13	12.8%	16.5%	14.4%	8.5%	22.0%	12.3%	6.6%	25.6%	10.5%
SJTULTLAB	14	9.6%	12.4%	10.8%	7.8%	20.2%	11.3%	6.2%	24.0%	9.9%
LIKEY	15	11.6%	15.0%	13.1%	7.9%	20.4%	11.4%	5.9%	22.7%	9.3%
UvT	16	11.4%	14.7%	12.9%	7.6%	19.6%	11.0%	5.8%	22.5%	9.2%
UNICE	17	8.8%	11.4%	9.9%	6.4%	16.5%	9.2%	5.5%	21.5%	8.8%
POLYU	18	3.8%	4.9%	4.3%	4.1%	10.6%	5.9%	4.1%	16.0%	6.6%
UKP	19	1.6%	2.1%	1.8%	0.9%	2.3%	1.3%	0.8%	3.1%	1.3%

Table: Ranked by F_1@15

Rankings and F-score per ACM category on combined keywords

Rank	C (Distr. Systems)	H (IR)	I (AI)	J (Social Sci.)
1	HUMB(28.3%)	HUMB(30.2%)	HUMB(24.2%)	HUMB(27.4%)
2	ICL(27.2%)	WINGNUS(28.9%)	SEERLAB(24.2%)	WINGNUS(25.4%)
3	KP-Miner(25.5%)	SEERLAB(27.8%)	KP-Miner(22.8%)	ICL(25.4%)
4	SZTERGAK(25.3%)	KP-Miner(27.6%)	KX_FBK(22.8%)	SZTERGAK(25.17%)
5	WINGNUS(24.2%)	SZTERGAK(27.6%)	WINGNUS(22.3%)	KP-Miner(24.9%)
6	KX_FBK(24.2%)	ICL(25.5%)	SZTERGAK(22.25%)	KX_FBK(24.6%)
7	DERIUNLP(23.6%)	KX_FBK(23.9%)	ICL(21.4%)	UNICE(23.5%)
8	SEERLAB(22.0%)	MAUI(23.9%)	DERIUNLP(20.1%)	SEERLAB(23.3%)
9	DFKI(21.7%)	DERIUNLP(23.6%)	DFKI(19.3%)	DFKI(22.2%)
10	MAUI(19.3%)	UNPMC(22.6%)	BUAP(18.5%)	MAUI(21.3%)
11	BUAP(18.5%)	SJTULTLAB(22.1%)	SJTULTLAB(17.9%)	DERIUNLP(20.3%)
12	JU_CSE(18.2%)	UNICE(21.8%)	JU_CSE(17.9%)	BUAP(19.7%)
13	LIKEY(18.2%)	DFKI(20.5%)	MAUI(17.6%)	JU_CSE(18.6%)
14	SJTULTLAB(17.7%)	BUAP(20.2%)	UNPMC(17.6%)	UNPMC(17.8%)
15	UvT(15.8%)	UvT(20.2%)	UNICE(14.7%)	LIKEY(17.2%)
16	UNPMC(15.2%)	LIKEY(19.4%)	LIKEY(11.3%)	SJTULTLAB(16.7%)
17	UNIC(14.3%)	JU_CSE(17.3%)	POLYU(13.6%)	POLYU(14.3%)
18	POLYU(12.5%)	POLYU(15.8%)	UvT(10.3%)	UvT(12.6%)
19	UKP(4.4%)	UKP(5.0%)	UKP(5.4%)	UKP(6.8%)

Rankings and F-score per ACM category on reader keywords

Rank	C (Distr. Systems)	H (IR)	I (AI)	J (Social Sci.)
1	ICL(23.3%)	HUMB(25.0%)	HUMB(21.7%)	HUMB(24.7%)
2	KX_FBK(23.3%)	WINGNUS(23.5%)	KX_FBK(21.4%)	WINGNUS(24.4%)
3	HUMB(22.7%)	SEERLAB(23.2%)	SEERLAB(21.1%)	SZTERGAK(24.4%)
4	SZTERGAK(22.7%)	KP-Miner(22.4%)	WINGNUS(19.9%)	KX_FBK(24.4%)
5	DERIUNLP(21.5%)	SZTERGAK(21.8%)	KP-Miner(19.6%)	UNICE(23.8%)
6	KP-Miner(21.2%)	KX_FBK(21.2%)	SZTERGAK(19.6%)	ICL(23.5%)
7	WINGNUS(20.0%)	ICL(20.1%)	ICL(19.6%)	KP-Miner(22.6%)
8	SEERLAB(19.4%)	DERIUNLP(20.1%)	DFKI(18.5%)	SEERLAB(22.0%)
9	DFKI(19.4%)	DFKI(19.5%)	SJTULTLAB(17.6%)	DFKI(21.7%)
10	JU_CSE(17.0%)	SJTULTLAB(19.5%)	DERIUNLP(17.3%)	BUAP(19.6%)
11	Likey(16.4%)	UNICE(19.2%)	JU_CSE(16.7%)	DERIUNLP(19.0%)
12	SJTULTLAB(15.8%)	Maui(18.1%)	BUAP(16.4%)	Maui(17.8%)
13	BUAP(15.5%)	UNPMC(18.1%)	UNPMC(16.1%)	JU_CSE(17.9%)
14	Maui(15.2%)	Likey(16.9%)	Maui(14.9%)	Likey(17.5%)
15	UNICE(14.0%)	UvT(16.4%)	UNICE(14.0%)	UNPMC(16.6%)
16	UvT(14.0%)	POLYU(15.5%)	POLYU(11.9%)	SJTULTLAB(16.3%)
17	UNPMC(13.4%)	BUAP(14.9%)	Likey(10.4%)	POLYU(13.3%)
18	POLYU(12.5%)	JU_CSE(12.6%)	UvT(9.5%)	UvT(13.0%)
19	UKP(4.5%)	UKP(4.3%)	UKP(5.4%)	UKP(6.9%)

Discussion and Closing Remarks

- Upper-Bound Performance
  - Top systems return F_1 in the upper twenties
  - Theoretically, F-score of 89% is possible (given a max 81% recall & 100% precision)
  - Note: 100% precision impossible due to fixed thresholds employed
- Human upper bound performance: 33.6% (author-assigned keywords)
- Closing Remarks
  - Certainly state-of-the-art in keyphrase extraction
  - Still room for improvement