Evaluating N-gram based Evaluation Metrics for Automatic Keyphrase Extraction

Su Nam Kim, Timothy Baldwin, and Min-Yen Kan

COLING 2010

Overview

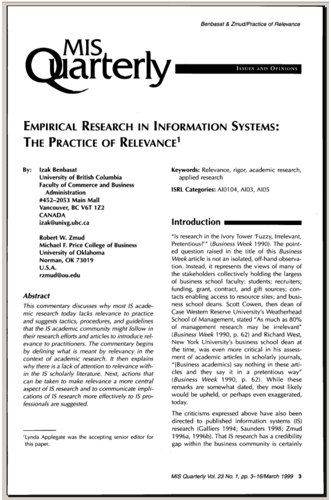
Goal: meta-evaluation of evaluation metrics for automatic keyphrase extraction

Keyphrases: phrases which capture the topic of an article

Significance: keyphrases used successfully in many NLP applications

- semantic metadata for summarization
- document indexing
- document clustering
- document summarization

Outline of Keyphrase Extraction



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

⇒ *relevance, rigor, academic research, applied research*

A Keyphrase Primer

- Keyphrases can be simplex words (e.g. *query* or *context-awareness*) or larger N-bars/noun phrases (e.g. *intrusion detection* or *mobile ad-hoc network*); the majority of keyphrases are 1–4 words long
- Keyphrases are normally composed of nouns and adjectives; they may contain hyphens (e.g. *multi-agent system*) and apostrophes (e.g. *Bayes' theorem*)
- Keyphrases can optionally incorporate PPs (e.g. *quality of service*); a variety of prepositions can be used (e.g. *incentive <u>for</u> cooperation*), but the genitive *of* is the most common
- Keyphrases can be coordinated (e.g. *performance and scalability*), and may also be abbreviations (e.g. *POMDP*)

Difficulties

Difficulties in Automatic Keyphrase Extraction Task

- Candidate selection: identify candidates, deal with lexical/constructional/semantic variations
- Candidate ranking: granularity/diversity/...
- Evaluation:
  - how to determine the appropriate number of machine-assigned keyphrases
  - how to treat lexical and semantic variations (i.e. near-misses)

Keyphrase Extraction Evaluation Metrics

- Standard approach to keyphrase extraction evaluation is based on Precision@N:
  - number of matching keywords in top-N
- Approaches for dealing with partial matches (lexical/constructional/semantic):
  - allow only pre-identified instances of constructional alternation (e.g. A of B $\rightarrow$ B A)
  - Semantic Similarity
    - use large-scale (domain-specific) corpora to estimate the semantic similarity between candidates, to support partial credit for candidates not in the gold standard
    - use link structure (e.g. in Wikipedia) to predict keyphrase equivalence
  - Domain Specific Thesaurus
    - use thesauri to check for term similarity using a thesaurus

Other Related Evaluation Metrics

- BLEU: measuring the relative similarity between a candidate translation and a set of reference translations based on *n*-gram composition
- METEOR: once again, calculate similarity based on string-level similarity, but include stem variation and WordNet synonymy
- NIST: once again, string-based, but weight up *n*-grams that occur less frequently, according to their information value
- ROUGE: based on *n*-gram overlap between candidate and reference summaries (or translations), with variations using co-occurrence statistics (ROUGE-N) or longest common subsequence (LCS)-based statistics (ROUGE-L)

R-precision

- *N*-gram based evaluation metric for automatic keyphrase extraction
- Treats near-misses by considering partial matches
- Three types of near-misses:
  - *INCLUDE*: *topic importance* vs. *topic*
  - *PARTOF*: *scheduling* vs. *real-time scheduling*
  - *MORPH*: *performance metric* vs. *performance metrics*

$$\text{R-precision} = \frac{\textit{number of overlapping segments}}{\textit{length of keyphrase}}$$

Modified R-precision (Our Proposed Metric) (I)

- Partial matching: give credit to partial matches according to their relative position in the candidate (e.g. *grid computing* for *grid computing algorithm*)
  - the closer to the head noun, the higher the weight: *fast computing system* $\rightarrow$ *fast* $<$ *computing* $<$ *system*
- Component weight: weight each component word w.r.t. their relative location in the keyphrase:

$$\text{CW} = \frac{1}{N - i + 1} \; (\textit{from left}, \; i = 1..N)$$

$$\text{Mod. R-precision} = \frac{\sum \text{CW in substring}}{\sum \text{CW in keyphrase}} (\times \; \textit{Frequency Weight})$$

Modified R-precision (Our Proposed Metric) (II)

- Example: *AB* from *ABC* = $\frac{\frac{1}{3}+\frac{1}{2}}{\frac{1}{3}+\frac{1}{2}+\frac{1}{1}} = \frac{5}{11}$
- Relative to gold-standard keyphrase *effective grid computing algorithm*:
  computing algorithm $>$ *grid computing* $>$ *effective grid*

Gold-Standard Keyphrases

- Compiled collection of 250 papers across 4 different categories from the ACM Digital Library
- Assigned reader-assigned keyphrases by hiring 50 human annotators, in addition to extracting the author-assigned keyphrases

	Author	Reader	Total
Total	1298/1305	3110/3221	3816/3962
NP/Nbars	937	2537	3027
Average	3.85/4.01	12.44/12.88	15.26/15.85
Found	769	2509	2864

Keyphrase Candidate Extraction

- Converted each PDF to text, POS-tagged/lemmatised the texts, and extracted keyphrase candidates via:
  - (**Rule1**) Nbar = `(NN*|JJ*)*(NN*)`
    e.g. *complexity, effective algorithm*, *distributed web-service discovery architecture*
  - (**Rule2**) Nbar `IN` Nbar
    e.g. *quality of service, sensitivity of VOIP traffic*, *simplified instantiation of zebroid*
- Excluded all simplex candidates with frequency of 1

Analysis of Human Assigned Scores

- Hired 4 human annotators to score semantic similarity between candidates and gold-standard keyphrases
- Scores: $[0, 4]$
- Broken down into three categories:
  - Head: candidate contains the head noun
  - First: candidate contains the first word of the keyphrase
  - Middle: neither HeadS nor FirstS

Evaluation Method: Correlation with Human Scores

- Comparison with human judgement:
  - annotators were given 3,248 keyphrase candidates
- Interpretation of human judgements:
  - average
  - majority
  - one-vs-rest inter-annotator correlation
- Comparator evaluation metrics:
  - BLEU, METEOR, NIST, ROUGE
- Evaluate each of the evaluation metrics via Spearman rank correlation

Rank Correlation between Human Majority and Machine Scores

		Human	R-precision Orig	R-precision Mod	BLEU	METEOR	NIST	ROUGE
Ave.	All	.4506	.4763	.2840	.3250	.3246	.3366	.3246
	$L \leq 4$	.4510	.5264	.2806	.3242	.3238	.3369	.3240
	$L \leq 3$	.4551	.4834	.2893	.3439	.3437	.3584	.3437
Maj.	All	.4603	.4763	.3438	.3407	.3403	.3514	.3404
	$L \leq 4$	.4604	.5264	.3434	.3423	.3421	.3547	.3422
	$L \leq 3$	.4638	.4838	.3547	.3679	.3675	.3820	.3676

Breakdown of Results (Average)

		Human	R-precision Orig	Mod	BLEU	METEOR	NIST	ROUGE
Loc	First	**.5508**	**.5032**	**.5033**	.3844	.3844	.4057	.3844
	Middle	**.5329**	**.5741**	**.5988**	**.4669**	**.4669**	.4055	**.4669**
	Head	**.3783**	**.4838**	**.4838**	.3865	.3860	.3780	.3864
Comp	Simple	.4452	**.4715**	.2790	.3653	.3445	.3527	.3445
	PP	**.4771**	**.4814**	.1484	.3367	.3122	.3443	.3123
	CC	.3645	.3810	.3140	.3748	.3446	.3384	.3748
POS	AdjN	**.4616**	**.4844**	.3507	.3147	.3132	.3115	.3133
	NN	.4467	**.4586**	.2581	.3321	.3321	.3488	.3322

Breakdown of Results (Majority)

		Human	R-precision Orig	R-precision Mod	BLEU	METEOR	NIST	ROUGE
Loc	First	**.5642**	**.5162**	**.5163**	.4032	.4032	.4297	.4032
	Middle	**.5510**	**.4991**	**.5320**	.4175	.4175	.3653	.4175
	Head	.4147	**.5073**	**.5074**	.4156	.4153	.4042	.4156
Comp	Simple	.4580	**.4869**	.3394	.3653	.3651	.3715	.3651
	PP	**.4715**	**.5068**	.3724	.3367	.3367	.3652	.3367
	CC	**.5777**	**.5513**	.3841	**.5745**	**.5571**	**.5600**	**.5745**
POS	AdjN	.4501	**.4861**	.3968	.3266	.3251	.3246	.3252
	NN	**.4631**	**.4733**	.3244	.3499	.3499	.3648	.3500

Findings

- Overall, R-precision achieved the highest correlation with humans (above inter-annotator agreement)
- Relatively little difference between *n*-gram-based evaluation metrics
- Correlation increases with the length of the (gold-standard) keyphrase
- modified R-precision superior to R-precision when we break down the results according to match position, but otherwise inferior (esp. over keyphrases including prepositions)

Conclusion

- Carried out meta-evaluation of keyphrase evaluation metrics
- Proposed a modification to R-precision, incorporating weighting of component words
- Compared keyphrase evaluation metrics to MT/summarisation evaluation metrics, and established that they are (on the whole) superior
- Confirmed the utility of R-precision for keyphrase extraction evaluation