Re-examining Automatic Keyphrase Extraction Approaches in Scientific Articles

Su Nam Kim and Min-Yen Kan

The University of Melbourne
National University of Singapore

Multiword Expression Workshop
August 6th 2009

Overview (I)

Goal: automatically extract keyphrases to represent the topic of articles

Keyphrases: Words which represent the topic of articles

Difficulties:

- identify term vs. non-term (*candidate selection*)
- dealing with variations (*candidate selection*)
- specification vs. generalization (*ranking candidates*)

Overview (II)

Significance used for many NLP applications

- semantic metadata for summarization (Barzilay:1997, Lawrie:2001, DÁvanzo:2005)
- document indexing (Gutwin:1999)
- document clustering (Zhang:2004, Hammouda:2005)
- document summarization (Berger:2000, Buyukkokten:2001)

Outline

Related Work (I)

- KEA (Frank:1999, Witten:1999, Medelyan:2006)
  - TF * IDF, first occurrance of word
  - domain specific (index as candidates)
- GenEx (Turney:1999, 2000)
  - 9 different syntactic features such as length, frequency of stem etc., decision tree induction
- Textract (Park:2004)
  - domain-specific cohesion (Damerau:1993) & term cohesion (Dice:1945)
- (Barker:2000)
  - using length, frequency & head noun frequency

Related Work (II)

- Turney:2003 – Keyphrase cohesion (among top N and the remaining, check keyphrase cohesion)
- Tomokiyo:2003 – using information loss between foreground & background data based on 1 vs. n-gram models
- Nguyen:2007 – using lingistic features such as section, POS sequence
- Fung:1998 – automatic keyphrase extraction in Chinese and Japanese
- Wan:2008 – referring clustered documents as domain info

Nature of Keyphrases

- form: simplex nouns or noun phrases (NPs)
- NPs as keyphrases: nouns with adjective(s), occasionally adverbs or other POSs (e.g. *dynamically allocated task*)
- can contain hypens (e.g. *sensor-grouping, multi-agent system*) and apostrophes (e.g. *Bayes' theroem, agent's goal*)
- length observation: few 3-term noun sequences are longer than 3-term NPs (Paukkeri:2008)
- many forms contain prepositions (e.g. *quality <u>of</u> service, incentive <u>for</u> cooperation*)
- few forms in conjunctive form (e.g. *behavioral and evolution and extrapolation*)
- can occur as abbreviations (e.g. *POMDP = partially observable Markov decision process*)

Keyphrase Variation

- *word order* fixed (e.g. *service quality* ≠ *quality service*)
- *word adjacency* fixed (e.g. *quality serivce* ≠ *quality ... service*)
- *morphological* variation allowed (e.g. *quality/qualities/...*)
- *lexical semantics* allowed, but costly to check (e.g. *multiagent behavior = multiagent action/manner*)
- *string overlap* allowed (e.g. *grid computing = grid computing algorithm*)

Candidate Selection: Approaches

- Issues: length, frequency, form, variation
- Aim: generalization with maximum coverage
- KEA uses the index words as candidates (Food & Agriculture domain)
- GenEx uses $1-3$ sequence words
- Textract uses regular expressions to extract noun sequences
- Nguyen & Kan uses regular expressions to extract both noun sequences and simple NP w/ preposition, *of* (i.e. *NN of NN*)

Candidate Selection: Proposed

Rule
(**Rule1**) *Frequency heuristic* freq $\geq$ 2 for simplex words vs. freq $\geq$ 1 for NPs
(**Rule2**) *Length heuristic* up to length 3 for NPs in non-*of-PP form* vs. up to length 4 for NPs in *of-PP form* (*synchronous concurrent program vs. model of multiagent interaction*)
(**Rule3**) *of-PP form alternation* (e.g. *number of sensor = sensor number,* *history of past encounter = past encounter history*) (**Rule4**) *Possessive alternation* (*agent's goal = goal of agent, security's value = value of security*)
(**Rule5**) *Noun Phrase* = $(NN\|NNS\|NNP\|NNPS\|JJ\|JJR\|JJS)*(NN\|NNS\|NNP\|NNPS)$ (*complexity, effective algorithm, grid computing, distributed discovery architecture*) (**Rule6**) *Noun Phrase* <u>*IN*</u> *Noun Phrase* (*quality of service, sensitivity of VOIP traffic,* ***VOIP traffic***, *simplified instantiation of zebroid,* ***simplified instantiation***)

Feature Engineering

1. Document Cohesion: How likely keyphrases are correlated with the document
2. Keyphrase Cohesion: Whether keyphrases share the same or similar semantics
3. Term Cohesion: High if the components make up a likely keyphrases (Church & Hanks 1989)
4. Other features

Use convention of "U"("S") to denote features more suited for (un)supervised approaches.
"*" also marks modified features not directly reported in previous work.

1. Document Cohesion (I)

- F1: TF*IDF *(S,U)* Frank:1999, Witten:1999
  - (F1a) *TF*IDF*
  - (F1b*) *TF* including counts of substrings
  - (F1c*) *TF* of substring as a separate feature
  - (F1d*) normalized *TF* by candidate types (i.e. simplex words vs. NPs)
  - (F1e*) normalized *TF* by candidate types as a separate feature
  - (F1f*) *IDF* using GOOGLE N-GRAM

1. Document Cohesion (II)

- F2: First Occurrence *(S,U)* Frank:1999, Witten:1999
- F3: Section Information *(S,U)* Nguyen:2007, abstract, introduction, conclusion, section head, title and/or references
- F4*: Additional Section Information
  - (F4a*) section, 'related/previous work'
  - (F4b*) counting substring occurring in key sections
  - (F4c*) section *TF* across all key sections
  - (F4d*) weighting key sections according to the portion of keyphrases found
- F5*: Last Occurrence *(S,U)*

2. Keyphase Cohesion

- F6*: Co-occurrence of Another Candidate in Section *(S,U)*
- F7*: Title overlap *(S)*
  - (F7a*) co-occurrence (Boolean) in title collocation
  - (F7b*) co-occurrence (*TF*) in title collection
- F8: Keyphrase Cohesion *(S,U)* Turney:2003

3. Term Cohesion

- F9: Term Cohesion *(S,U)*
  - (F9a) term cohesion by Park:2004
  - (F9b*) normalized *TF* by candidate types (i.e. simplex words vs. NPs)
  - (F9c*) applying different weight by candidate types
  - (F9d*) normalized *TF* and different weighting by candidate types

4. Other Features

- F10: Acronym *(S)* Nguyen:2007
- F11: POS sequence *(S)* Hulth:2006
- F12: Suffix sequence *(S)* Nguyen:2007
- F13: Length of Keyphrases *(S,U)* Barker:2000

Evaluation

- Exact Matching Scheme:
  number of matching keywords in top N_{th}
  - partial matching doesn't receive credits
  - very limited variation of keyphrases (e.g. A of B $->$ B A)
- Semantic Similarity (Jamasz:2004)
  - using terabyte corpus to measure the Top candidates and keyphrases
  - require large corpus to measure it
- Domain Specific Thesaurus (Medelyan:2006)
  - using Agrovoc (thesaurus: food & agriculture), check similar words
- Wikipedia InterLink (Paukkeri:2008)
  - using the interlink among the multilingual documents

Experimental Setup

- Targets
  1. test proposed candidate selection & length heuristics & alternation
  2. test features over supervised vs. unsupervised approaches
- Simulated Systems
  - KEA: (Frank:1999,Witten:1999) *(S,U)*, TF*IDF(F1), First occurrence(F2)
  - N&K: (Nguyen:2007) *(S)*, TF*IDF(F1), First occurrence(F2), Section information(F3), Acronym(F10), POS sequence(F11), Suffix sequence(F12)
  - Baseline: modified N&K. remove Acronym(F10), POS sequence(F11), Suffix sequence(F12) and add additional section information(F4a) *(S,U)*

Results: Candidate Selection

Method	Features	C	Fifteen Match	Fifteen Precision	Fifteen Recall	Fifteen F-score
All Candidates	KEA	U	0.13	0.88%	0.86%	0.87%
		S	1.84	12.24%	12.03%	12.13%
	N&K	S	2.54	16.93%	16.64%	16.78%
	baseline	U	2.20	14.64%	14.39%	14.51%
		S	2.44	16.24%	15.96%	16.10%
Length<=3 Candidates	KEA	U	0.13	0.88%	0.86%	0.87%
		S	1.84	12.24%	12.03%	12.13%
	N&K	S	2.62	17.49%	17.19%	17.34%
	baseline	U	2.20	14.64%	14.39%	14.51%
		S	2.40	16.00%	15.72%	15.86%
Length<=3 Candidates + Alternation	KEA	U	0.07	0.48%	0.47%	0.47%
		S	1.87	12.45%	12.24%	12.34%
	N&K	S	2.88	19.20%	18.87%	19.03%
	baseline	U	2.37	15.79%	15.51%	15.65%
		S	2.69	17.92%	17.61%	17.76%

Table: Performance on Proposed Candidate Selection

Results: Feature Engineering

- Best Features: *F1c:TF of substring as a separate feature*, *F2:first occurrence*, *F3:section information*, *F4d:weighting key sections*, *F5:last occurrence*, *F6:co-occurrence of another candidate in section*, *F7b:title overlap*, *F9a:term cohesion by (Park:2004)*, *F13:length of keyphrases*

Features	C	Fifteen			
		Match	Prec.	Recall	F-score
Best	U	*2.61*	.174	.171	.173
	S	*3.15*	.210	.206	.208
Best w/o TF*IDF	U	*2.61*	.174	.171	.173
	S	*3.12*	.208	.204	.206

Table: Performance on Feature Engineering

Experiment (IV)

A	Method	Feature
+	S	F1a,F2,F3,F4a,F4d,F9a
	U	F1a,F1c,F2,F3,F4a,F4d,F5,F7b,F9a
-	S	F1b,F1c,F1d,F1f,F4b,F4c,F7a,F7b,F9b-d,F13
	U	F1d,F1e,F1f,F4b,F4c,F6,F7a,F9b-d
?	S	F1e,F10,F11,F12
	U	F1b

Table: Performance on Each Feature

Conclusion

We have explored two issues for MWE in scientific articles:

1. **Candidate Selection**
   - subject to the performances directly
   - maximum coverage as well as standard method needed
   - explored heuristics (i.e. length, frequency, alternation)
2. **Feature Engineering**
   - tested features w.r.t. supervised vs. unsupervised approaches
   - steady study but need to be improved for NLP applications, especially unsupervised approaches

SemEval 2: Keyphrase Extraction in Scientific Documents

- 2010 will feature a shared task on keyphrase extraction
- We look forward to your participation on the task!
- http://semeval2.fbk.eu/semeval2.php?location=tasks