SYSTEMS ENGINEERING RESEARCH CENTER

SE Business & Analytics

Sponsor: CCDC

By

Dr. K.P. (Suba) Subbalakshmi

Founding Director, Stevens Institute for Artificial Intelligence

Prof. Dept of Electrical and Computer Engineering

11th Annual SERC Sponsor Research Review

November 19, 2019

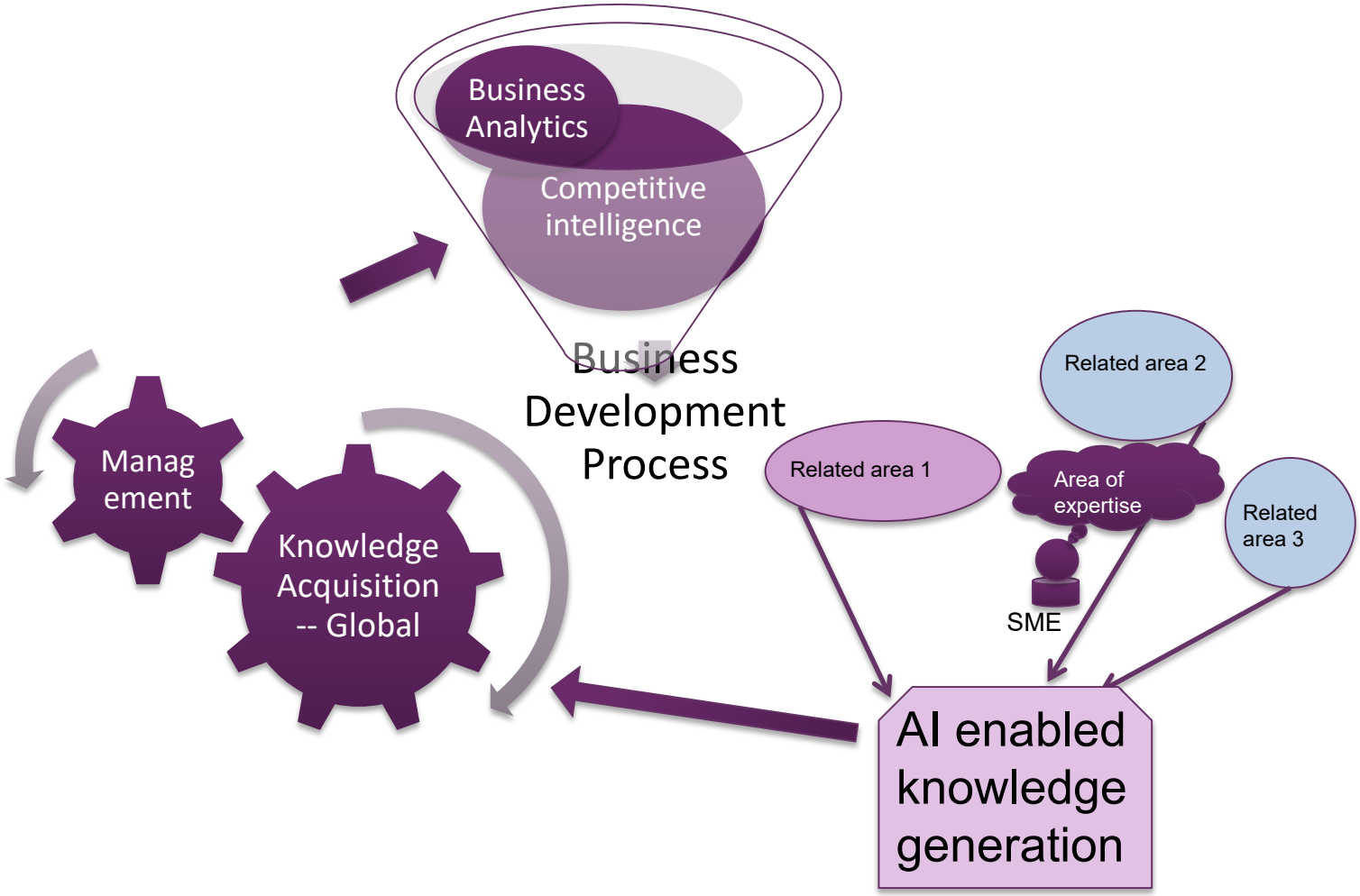
FHI 360 CONFERENCE CENTER

1825 Connecticut Avenue NW, 8th Floor

Washington, DC 20009

www.sercuarc.org

RT-213: SE Business & Analytics



SYSTEMS ENGINEERING RESEARCH CENTER

AI Enabled Keyphrase Extraction

- Keyphrase extraction is the first step for a lot of downstream NLP tasks needed for competitive intelligence
  - Select relevant articles for SME's attention
  - Rank ideas according to importance …
  - Extract trends
  - Discover areas of interest

Project Steps

- Use sponsor provided keyphrases to gather relevant datasets
- Clean dataset
- Implement basic keyphrase extraction algorithms
- Propose improvements on these
- Test extensively on bespoke and standard datasets
- Keyphrases provided by CCDC

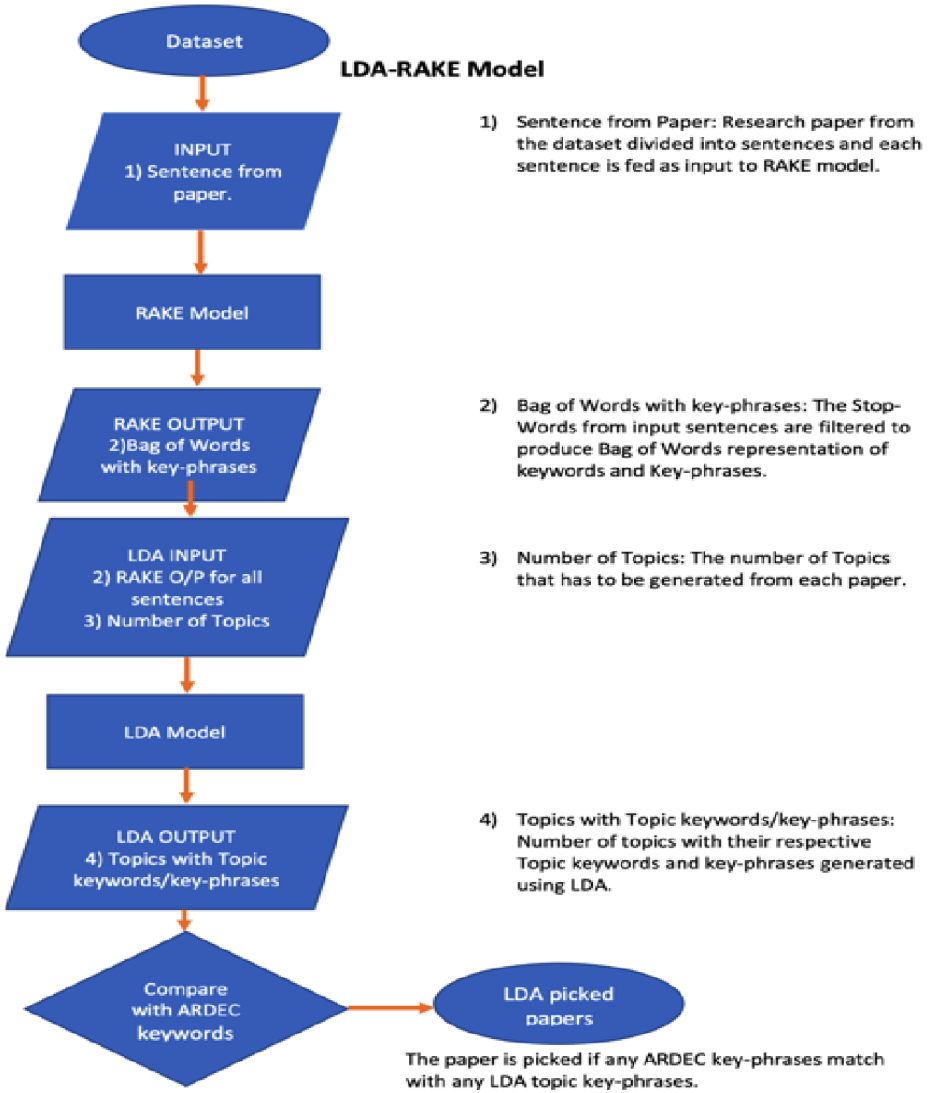
Additive technology
Chemical additive to solutions
Cryogenic milling
Microfluidics
Nanopowder
Nanoscale

Datasets

Dataset	Articles	Number of Papers
CCDC	Propellants_Explosives_Pyrotechnics2010, Propellants_Explosives_Pyrotechnics2011, Propellants_Explosives_Pyrotechnics2013, Propellants_Explosives_Pyrotechnics2014, Propellants_Explosives_Pyrotechnics2015, Propellants_Explosives_Pyrotechnics2016, Propellants_Explosives_Pyrotechnics2017, Propellants_Explosives_Pyrotechnics2018	684 → 139
ACS Dataset	ACS Environmental Science & Technology 2011 ACS Nano 2007 ACS Nano 2011 ACS Environmental Science & Technology 2012 70 Random papers selected from ACS journals targeting the keywords provided by CCDC	681
NUS Dataset	This is one of the original datasets that was used for Position Rank Analysis(PRA) [4]	215

- PDF → text conversion using tool PDFMiner
- Non-standard formatting of papers introduced lot of noise in the form of special characters
- Removed using python filters.

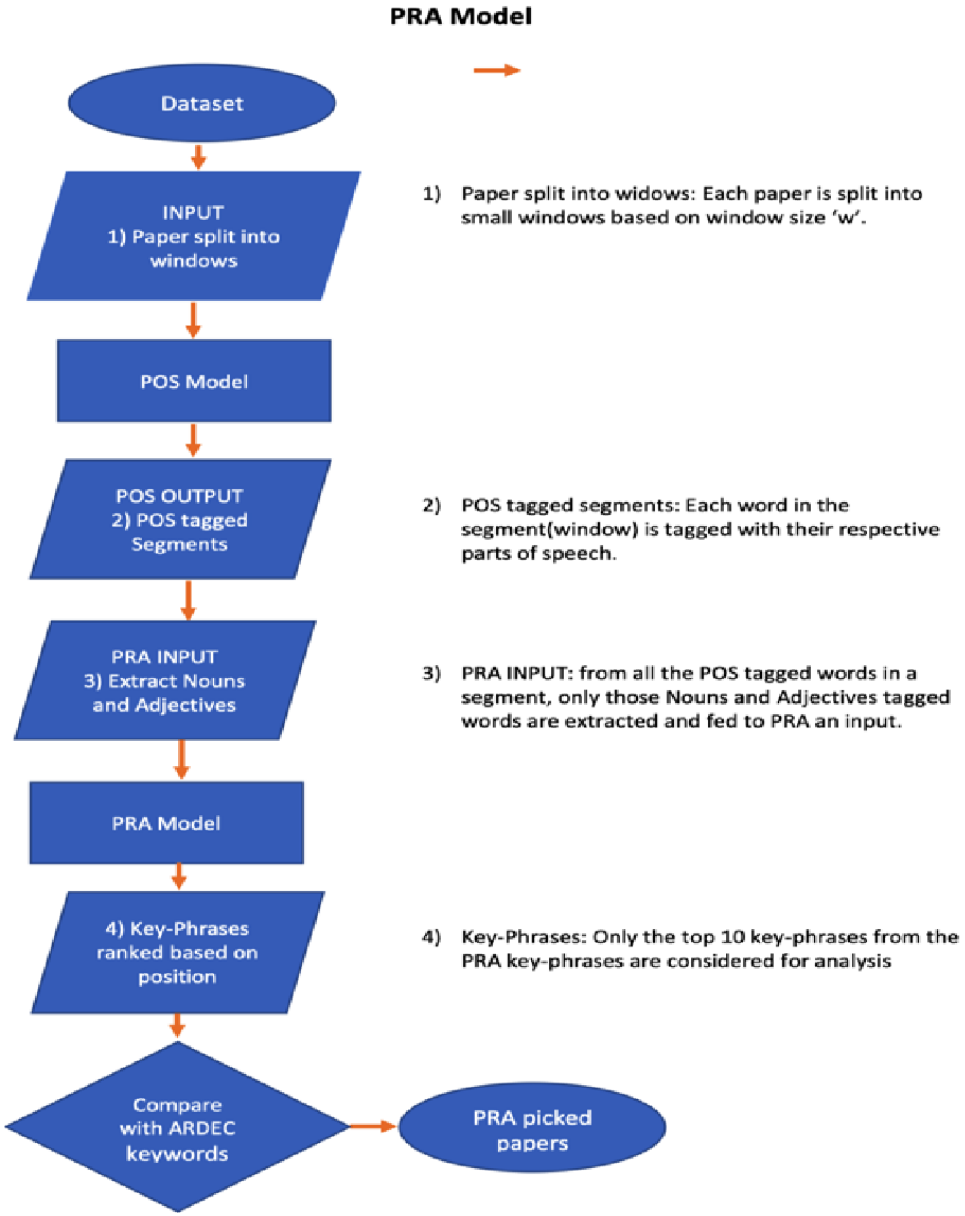
The Architectures: LDA-RAKE



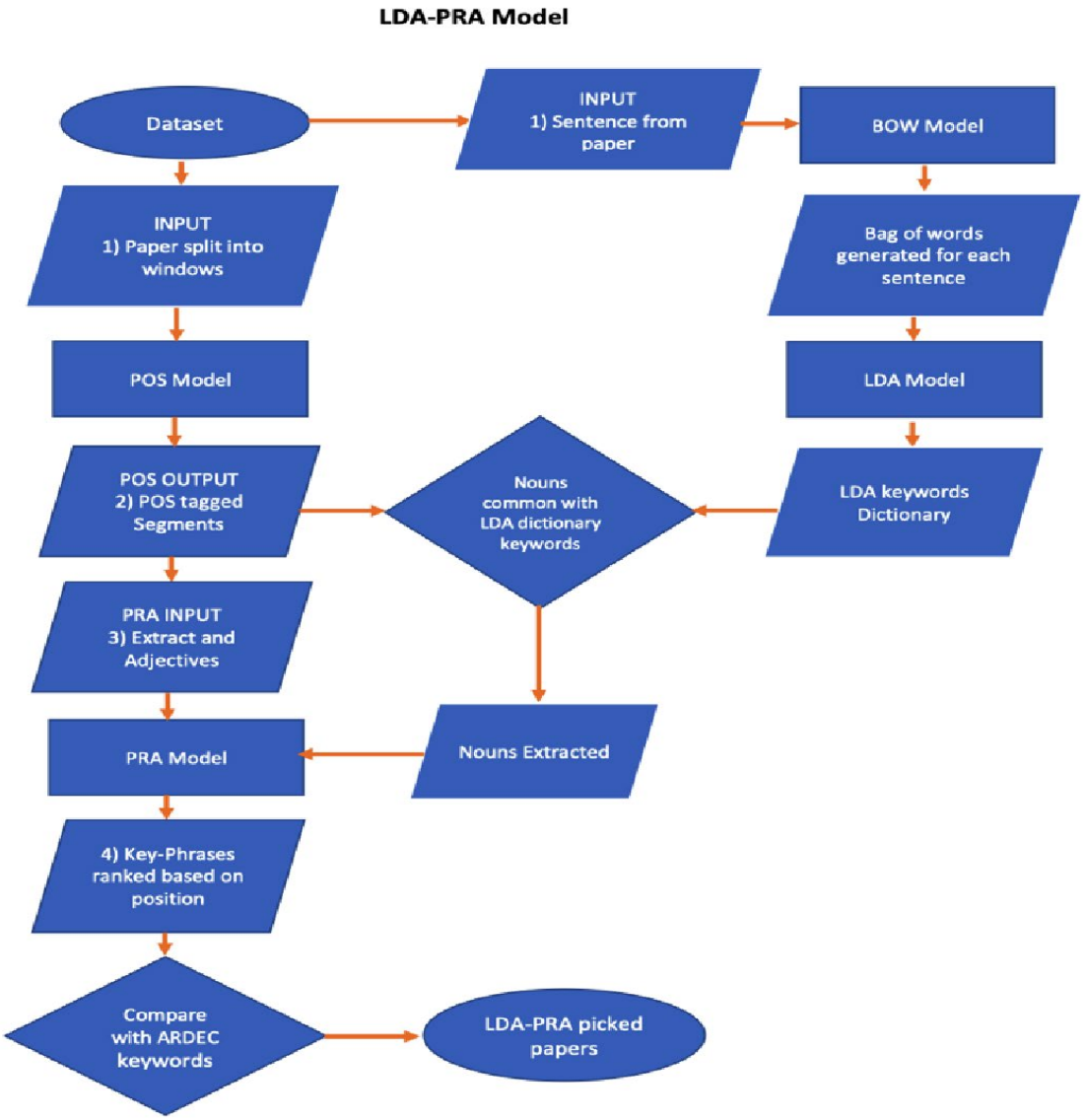
- Some quick notes:
  - LDA (Latent Dirichlet Allocation) is a topic modeler. Generates n topics, which are essentially probability density functions
  - RAKE (Rapid Automated Keyword Extraction) is a keyword extractor, but works only on a sentence to sentence basis.

The Architectures: PRA

- Position Rank Analysis (PRA)
  - —unsupervised frequency-based keyphrase extractor based on ranks from a text corpus
  - — co-occurrence score of keywords is calculated based on position of sentence in document
    - Earlier => higher
  - —Candidate keyphrases are consecutive adjective noun pairs



The Architectures: LDA-PRA



- LDA-PRA
  - Uses LDA picked keywords as the nouns instead of just any noun in the LDA-PRA architecture
  - Should allow more customization for the paper at hand

SYSTEMS ENGINEERING RESEARCH CENTER

Reducing SME Workload

Number of LDA topics	LDA-RAKE (for ACS dataset 681 papers)			
	Avg Precision	Average Recall	Avg F-Score	Total # papers picked
10	0.1	0.166	0.125	19
15	0.1	0.166	0.125	19
20	0.1	0.166	0.125	13

- Extracting articles of interest based on SME's general description keyphrase
- LDA-RAKE extracted 6 papers out of CCDC dataset

SYSTEMS ENGINEERING RESEARCH CENTER

Results: NUS Dataset

# LDA topics20	Only-PRA (# LDA topics not relevant here)				LDA-PRA			
	Avg. Precision	Average Recall	Average F-Score	# papers with matching keyphrases	Avg Precision	Avg Recall	Average F-Score	# papers with matching keyphrases
10	0.2211	0.2328	0.2060	180	0.2012	0.2117	0.1895	161
15	0.2211	0.2328	0.2060	180	0.2061	0.2161	0.1934	162
20	0.2211	0.2328	0.2060	180	0.2068	0.2150	0.1932	160

PRA does better on all metrics

SYSTEMS ENGINEERING RESEARCH CENTER

Results: CCDC Dataset – CCDC provided keyphrases

# LDA topics	Only-PRA (# LDA topics not relevant here)				LDA-PRA			
	Avg. Precision	Average Recall	Average F-Score	# papers with matching keyphrases	Avg Precision	Avg Recall	Average F-Score	# papers with matching keyphrases
10	0.1	0.166	0.125	3	0.1	0.166	0.125	4
15	0.1	0.166	0.125	3	0.1	0.166	0.125	3
20	0.1	0.166	0.125	3	0.1	0.166	0.125	4

LDA-PRA does better in terms of number of papers it identifies as relevant

SYSTEMS ENGINEERING RESEARCH CENTER

Results: CCDC Dataset (139 papers) – Author provided keyphrases

# LDA topics	Only-PRA (# LDA topics not relevant here)				LDA-PRA			
	Avg. Precision	Average Recall	Average F-Score	# papers with matching keyphrases	Avg Precision	Avg Recall	Average F-Score	# papers with matching keyphrases
10	0.1632	0.4248	0.2291	87	0.1395	0.3289	0.1914	48
15	0.1632	0.4248	0.2291	87	0.1395	0.3348	0.1925	48
20	0.1632	0.4248	0.2291	87	0.1387	0.3273	0.1909	49

Note that only 139 papers had author provided keyphrases in this dataset

Conclusions

- Bespoke datasets were constructed for the sponsor
- Several keyphrase extraction algorithms were tried and proposed
- SME workload is cut, by using these keyphrase extraction algorithms to match the right paper for the SME's reading consideration
- Extensive experimentation shows that:
  - PRA works better in finding keyphrases that match the author provided keyphrases
  - PRA and LDA-PRA have mixed results with the bespoke dataset when it comes to finding papers that are of interested to any set of keyphrases provided
  - Better optimization of LDA parameters could change the result trend