Pattern Mining Applications: Mining Quality Phrases from Text Data

Pattern Mining Applications: Mining Quality Phrases from Text Data

- From Frequent Pattern Mining to Phrase Mining
- Previous Phrase Mining Methods
- ToPMine: Phrase Mining without Training Data
- SegPhrase: Phrase Mining with Tiny Training Sets

Thanks to Ahmed El-Kishky@UIUC, Jialu Liu@UIUC, Jingbo Shang@UIUC, Xiang Ren@UIUC, Chi Wang@MSR and Marina Danilevsky@IBM for their contributions

From Frequent Pattern Mining to Phrase Mining

Why Phrase Mining?

- Unigrams vs. phrases
  - **Unigrams** (single words) are often *ambiguous*
    - Example: “United”: United States? United Airline? United Parcel Service?
  - **Phrase**: A natural, meaningful, *unambiguous* semantic unit
    - Example: “United States” vs. “United Airline”
- Mining semantically meaningful phrases
  - Transform text data from *word granularity* to *phrase granularity*
  - Enhance the power and efficiency at manipulating unstructured data

From Frequent Pattern Mining to Phrase Mining

- General principle
  - Exploit information redundancy and data-driven criteria to determine phrase boundaries and salience
- Methodology: Exploring three ideas
  - Frequent pattern mining and colocation analysis
  - Phrasal segmentation
  - Quality phrase assessment
- Recent developments of phrase mining methods
  - ToPMine: Mining quality phrase without training (A. El-Kishky, et al., 2015)
  - SegPhrase: Mining quality phrase with tiny training sets (J. Liu, et al., 2015)

Previous Phrase Mining Methods

Phrase Mining: Can We Reduce Annotation Cost?

- Phrase mining: Originated from the NLP community—"Chunking"
  - Model it as a sequence labeling problem (B-NP, I-NP, O, …)
- Need annotation and training
  - Annotate hundreds of documents as training data
  - Train a supervised model based on part-of-speech features
- Recent trend:
  - Use distributional features based on web n-grams (Bergsma et al., 2010)
  - State-of-the-art performance: ~95% accuracy, ~88% phrase-level F-score
- Limitations
  - High annotation cost, not scalable to a new language, a new domain/genre
  - May not fit domain-specific, dynamic, emerging applications
    - Scientific domains, query logs, or social media (e.g., Yelp and Twitter data)

Unsupervised Phrase Mining and Topic Modeling

- Many studies of unsupervised phrase mining are linked with topic modeling
- Topic modeling
  - Represents documents by multiple topics in different proportions
    - Each topic is represented by a word distribution
  - Does not require any prior annotations or labeling of the documents
- Statistical topic modeling algorithms
  - The most common algorithm: LDA (Latent Dirichlet Allocation) [Blei, et al., 2003]
- Three strategies on phrase mining with topic modeling
  - Strategy 1: Generate bag-of-words → generate sequence of tokens
  - Strategy 2: Post bag-of-words model inference, visualize topics with n-grams
  - Strategy 3: Prior bag-of-words model inference, mine phrases and impose on the bag-of-words model

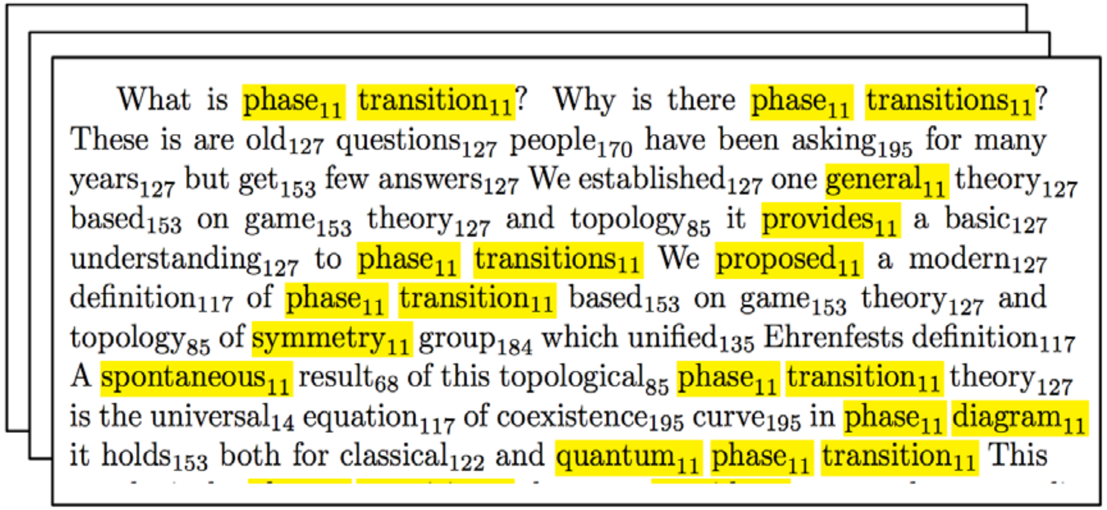
Strategy 1: Simultaneously Inferring Phrases and Topics

- Bigram Topic Model [Wallach'06]
  - Probabilistic generative model that conditions on previous word and topic when drawing next word
- Topical N-Grams (TNG) [Wang, et al.'07] (a generalization of Bigram Topic Model)
  - Probabilistic model that generates words in textual order
  - Create n-grams by concatenating successive bigrams
- Phrase-Discovering LDA (PDLDA) [Lindsey, et al.'12]
  - Viewing each sentence as a time-series of words, PDLDA posits that the generative parameter (topic) changes periodically
  - Each word is drawn based on previous m words (context) and current phrase topic
- Comments on this strategy
  - High model complexity: Tends to overfitting
  - High inference cost: Slow

Strategy 2: Post Topic-Modeling Phrase Construction (I): TurboTopics

- **TurboTopics** [Blei & Lafferty'09] – Phrase construction as a post-processing step to Latent Dirichlet Allocation
  - Perform Latent Dirichlet Allocation on corpus to assign each token a topic label
  - Merge adjacent unigrams with the same topic label by a distribution-free permutation test on arbitrary-length back-off model
  - End recursive merging when all significant adjacent unigrams have been merged

Annotated documents


What is phase$_{11}$ transition$_{11}$? Why is there phase$_{11}$ transitions$_{11}$? These is are old$_{127}$ questions$_{127}$ people$_{170}$ have been asking$_{195}$ for many years$_{127}$ but get$_{153}$ few answers$_{127}$ We established$_{127}$ one general$_{11}$ theory$_{127}$ based$_{153}$ on game$_{153}$ theory$_{127}$ and topology$_{85}$ it provides$_{11}$ a basic$_{127}$ understanding$_{127}$ to phase$_{11}$ transitions$_{11}$ We proposed$_{11}$ a modern$_{127}$ definition$_{117}$ of phase$_{11}$ transition$_{11}$ based$_{153}$ on game$_{153}$ theory$_{127}$ and topology$_{85}$ of symmetry$_{11}$ group$_{184}$ which unified$_{135}$ Ehrenfests definition$_{117}$ A spontaneous$_{11}$ result$_{68}$ of this topological$_{85}$ phase$_{11}$ transition$_{11}$ theory$_{127}$ is the universal$_{14}$ equation$_{117}$ of coexistence$_{195}$ curve$_{195}$ in phase$_{11}$ diagram$_{11}$ it holds$_{153}$ both for classical$_{122}$ and quantum$_{11}$ phase$_{11}$ transition$_{11}$ This

LDA topic #11

phase, transitions, phases, transition, quantum, critical, symmetry, field, point, model, order, diagram, systems, two, theory, system, study, breaking, spin, first

Turbo topic #11

phase transitions, model, symmetry, point, quantum, systems, phase transition, phase diagram, system, order, field, order, parameter, critical, two, transitions in, models, different, symmetry breaking, first order, phenomena

Post Topic-Modeling Phrase Construction (II): KERT

- **KERT** [Danilevsky et al.'14] – Phrase construction as a post-processing step to LDA
  - Run bag-of-words model inference and assign topic label to each token
  - Perform **frequent pattern mining** to extract candidate phrases within each topic
  - Perform **phrase ranking** based on four different criteria
    - **Popularity:** e.g., "information retrieval" vs. "cross-language information retrieval"
    - **Concordance**
      - "powerful tea" vs. "strong tea"
      - "active learning" vs. "learning classification"
    - **Informativeness:** e.g., "this paper" (frequent but not discriminative, not informative)
    - **Completeness:** e.g., "vector machine" vs. "support vector machine"

Comparability property: directly compare phrases of mixed lengths

ToPMine: Phrase Mining without Training Data

Strategy 3: First Phrase Mining then Topic Modeling

- Why first Phrase Mining then Topic Modeling?
  - With Strategy 2, tokens in the same phrase may be assigned to different topics
    - Ex. knowledge discovery using least squares support vector machine classifiers…
      - *Knowledge discovery* and *support vector machine* should have coherent topic labels
- Solution: switch the order of phrase mining and topic model inference

[knowledge discovery] using [least squares] [support vector machine] [classifiers] …	→	[knowledge discovery] using [least squares] [support vector machine] [classifiers] …

- Techniques for this strategy
  - Phrase mining, document segmentation, and phrase ranking
  - Topic model inference with phrase constraint

ToPMine: Phrase Mining before Topic Modeling

- **ToPMine** [El-Kishky et al. VLDB'15]: Phrase mining, then phrase-based topic modeling
- Phrase mining
  - Frequent ***contiguous pattern*** mining: Extract candidate phrases and their counts
  - Agglomerative merging of adjacent unigrams as guided by a **significance score**
  - Document segmentation to count phrase occurrence
    - Calculate rectified (i.e., true) phrase frequency
  - Phrase ranking (using the criteria proposed in KERT)
    - Popularity, concordance, informativeness, completeness
- Phrase-based topic modeling
  - The mined bag-of-phrases are passed as input to PhraseLDA, an extension of LDA, that constrains all words in a phrase to each sharing the same latent topic

Phrase	Raw frequency	Rectified frequency
[support vector machine]	90	80
[vector machine]	95	0
[support vector]	100	20

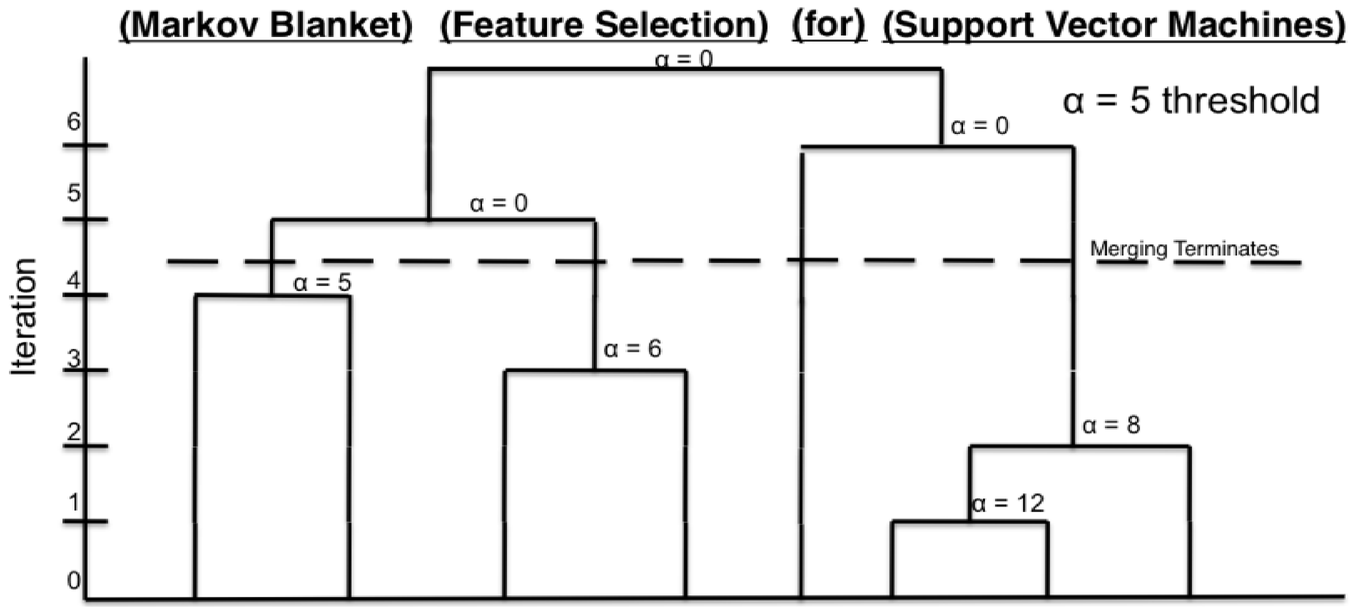
Collocation Mining

- Collocation: A sequence of words that occur more frequently than expected
  - Often “interesting”, relay information not portrayed by their constituent terms
    - Ex. “made an exception”, “strong tea”
- Many different measures used to extract collocations from a corpus [Dunning 93, Pederson 96]
  - E.g., mutual information, t-test, z-test, chi-squared test, likelihood ratio

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \qquad sig = \frac{count(phr_{x+y}) - E[count(phr_{x+y})]}{\sqrt{count(phr_{x+y})}} \qquad \chi^2 = \sum \frac{(O-E)^2}{E}$$

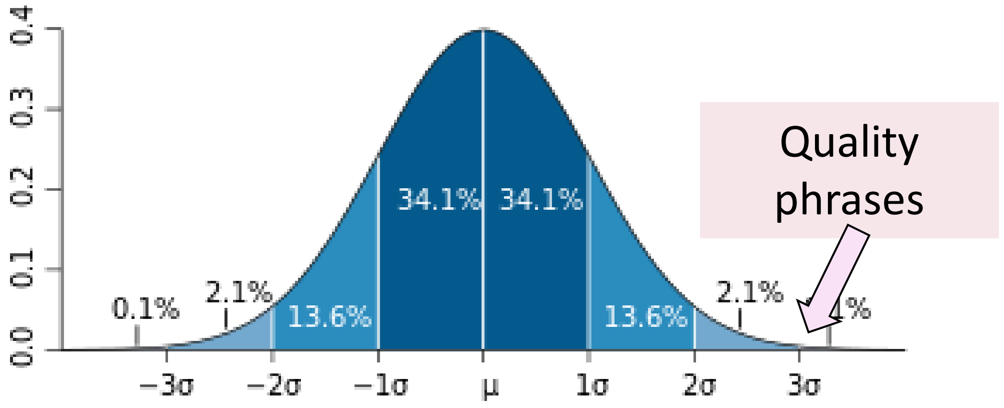
- Many of these measures can be used to guide the agglomerative **phrase-segmentation** algorithm

Phrase Candidate Generation: Frequent Pattern Mining + Statistical Analysis



Markov Blanket Feature Selection for Support Vector Machines.

[Markov blanket] [feature selection] for [support vector machines]
[knowledge discovery] using [least squares] [support vector machine] [classifiers]
…[support vector] for [machine learning]…



Based on significance score [Church et al.'91]:

$$\alpha(P_1, P_2) \approx (f(P_1 \bullet P_2) - \mu_0(P_1, P_2)) / \sqrt{f(P_1 \bullet P_2)}$$

Note for the first title:

- [feature selection] forms phrase but not [selection for] based on the significant scores computed
- [support vector machine] does not contribute to the counts of [support], [vector], [support vector], [vector machine]

ToPMine: Experiments on DBLP Abstracts

	Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*
unigrams	problem	word	data	programming	data
	algorithm	language	method	language	patterns
	optimal	text	algorithm	code	mining
	solution	speech	learning	type	rules
	search	system	clustering	object	set
	solve	recognition	classification	implementation	event
	constraints	character	based	system	time
	programming	translation	features	compiler	association
	heuristic	sentences	proposed	java	stream
	genetic	grammar	classifier	data	large
n-grams	genetic algorithm	natural language	data sets	programming language	data mining
	optimization problem	speech recognition	support vector machine	source code	data sets
	solve this problem	language model	learning algorithm	object oriented	data streams
	optimal solution	natural language processing	machine learning	type system	association rules
	evolutionary algorithm	machine translation	feature selection	data structure	data collection
	local search	recognition system	paper we propose	program execution	time series
	search space	context free grammars	clustering algorithm	run time	data analysis
	optimization algorithm	sign language	decision tree	code generation	mining algorithms
	search algorithm	recognition rate	proposed method	object oriented programming	spatio temporal
	objective function	character recognition	training data	java programs	frequent itemsets

ToPMine is efficient and generates high-quality topics and phrases without any training data

ToPMine: Experiments on Yelp Reviews

	Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*
unigrams	coffee	food	room	store	good
	ice	good	parking	shop	food
	cream	place	hotel	prices	place
	flavor	ordered	stay	find	burger
	egg	chicken	time	place	ordered
	chocolate	roll	nice	buy	fries
	breakfast	sushi	place	selection	chicken
	tea	restaurant	great	items	tacos
	cake	dish	area	love	cheese
	sweet	rice	pool	great	time
n-grams	ice cream	spring rolls	parking lot	grocery store	mexican food
	iced tea	food was good	front desk	great selection	chips and salsa
	french toast	fried rice	spring training	farmer's market	food was good
	hash browns	egg rolls	staying at the hotel	great prices	hot dog
	frozen yogurt	chinese food	dog park	parking lot	rice and beans
	eggs benedict	pad thai	room was clean	wal mart	sweet potato fries
	peanut butter	dim sum	pool area	shopping center	pretty good
	cup of coffee	thai food	great place	great place	carne asada
	iced coffee	pretty good	staff is friendly	prices are reasonable	mac and cheese
	scrambled eggs	lunch specials	free wifi	love this place	fish tacos

ToPMine works well for phrase and topic mining in social media data

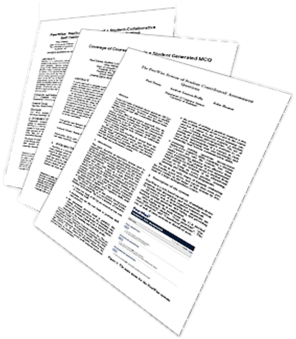
SegPhrase: Phrase Mining with Tiny Training Sets

SagPhrase: Phrase Mining with Tiny Training Sets

- A small set of training data may enhance the quality of phrase mining

J. Liu et al., Mining Quality Phrases from Massive Text Corpora. In *SIGMOD'*15

Raw Corpus



Quality Phrases

+ A small set of labels by human or a general KB

Segmented Corpus

Document 1
Citation recommendation is an interesting but challenging research problem in data mining area.

Document 2
In this study, we investigate the problem in the context of heterogeneous information networks using data mining technique.

Document 3
Principal Component Analysis is a linear dimensionality reduction technique commonly used in machine learning applications.

Input Raw Corpus → **Quality Phrases** ↔ **Segmented Corpus**

Phrase Mining **Phrasal Segmentation**

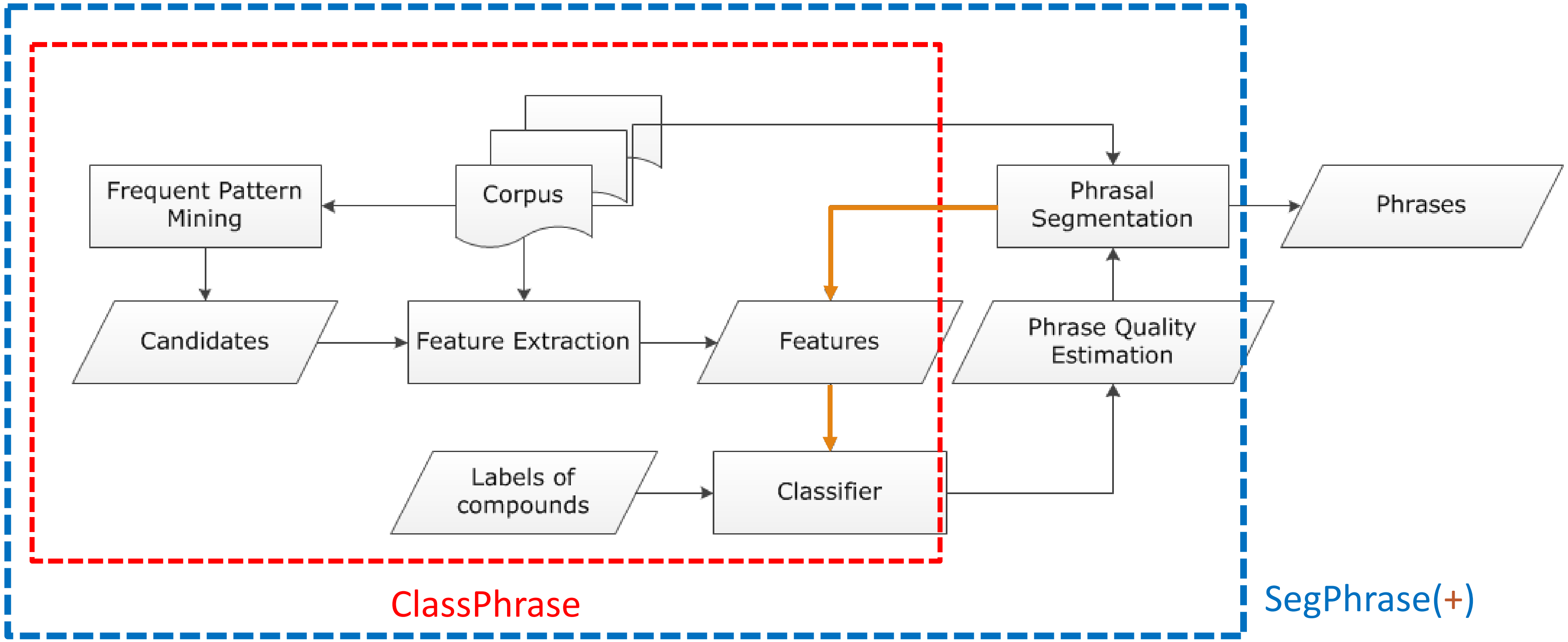
Integrating phrase mining with phrasal segmentation and classification

SegPhrase+: The Overall Framework

- ClassPhrase: Frequent pattern mining, feature extraction, classification
- SegPhrase: Phrasal segmentation and phrase quality estimation
- SegPhrase+: One more round to enhance mined phrase quality

SegPhrase (a classifier is used)

Small labeled dataset provided by experts
or
a distant supervised KB (e.g., Wikipedia / DBPedia)



SegPhrase: Pattern Mining and Feature Extraction

- **Pattern Mining for Candidate Set**
  - Build a candidate phrases set by frequent pattern mining
    - Mining frequent *k*-grams (*k* is typically small, e.g., 6 in the experiments)
    - **Popularity** measured by *raw* frequent words and phrases mined from the corpus
- **Feature Extraction: Concordance**
  - Partition a phrase into two parts to check whether the co-occurrence is significantly higher than pure random
- **Feature Extraction: Informativeness**
  - Quality phrases typically start and end with a non-stopword
    - “machine learning is” vs. “machine learning”
  - Use average IDF over words in the phrase to measure the semantics
  - Usually, the probabilities of a quality phrase in quotes, brackets, or connected by hyphen should be higher (punctuations information)
    - e.g., “state-of-the-art”

SegPhrase: Classification Using Tiny Training Sets

- Use tiny training sets (300 labels for 1GB corpus; can also use phrases extracted from KBs)
  - Label: indicating whether a phrase is a high quality one
    - E.g., “support vector machine”: 1; “the experiment shows”: 0
- Classification: Construct models to distinguish quality phrases from poor ones
  - Use *Random Forest* algorithm to bootstrap different datasets with limited labels
- Phrasal segmentation can tell which phrase is more appropriate
  - Ex: “A standard [feature vector] [machine learning] setup is used to describe .....”

    Not counted towards the rectified frequency
  - Partition a sequence of words by maximizing the likelihood
  - Consider length penalty and filter out phrases with low rectified frequency
- Process: Classification → Phrasal segmentation // SegPhrase

  → Classification → Phrasal segmentation // SegPhrase+

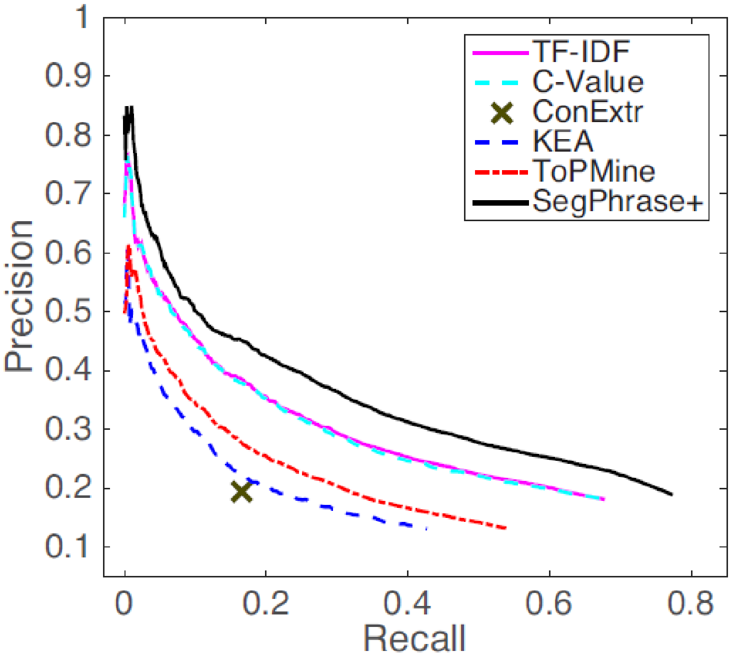
Performance: Precision Recall Curves on DBLP

- Datasets:

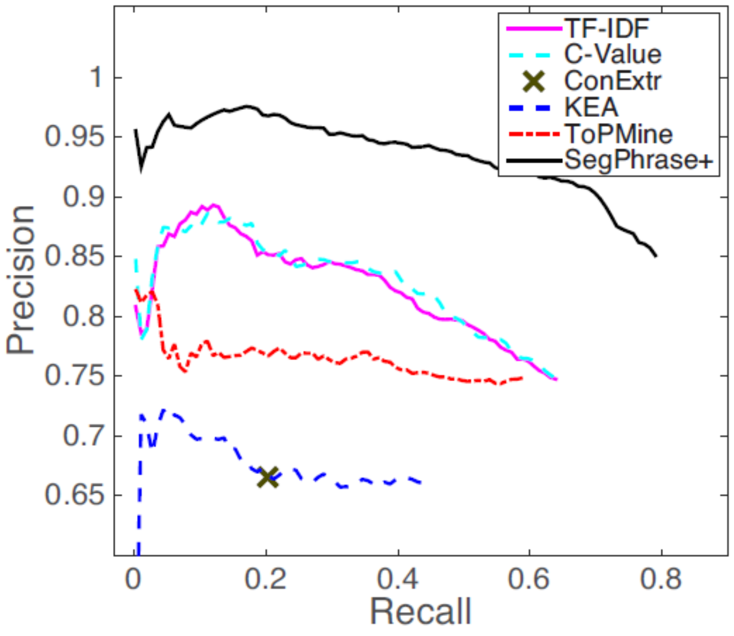
Dataset	#docs	#words	#labels
DBLP	2.77M	91.6M	300
Yelp	4.75M	145.1M	300

Use only 300 human labeled phrases for training

- Evaluation
  - Wiki Phrases (based on internal links, ~7K high quality phrases)
  - Sampled 500*7 Wiki-uncovered phrases: Results evaluated by 3 reviewers
- Compared with other phrase-mining methods
  - TF-IDF, C-Value, ConExtr, KEA, and ToPMine
- Also, Segphrase+ is efficient, linearly scalable



Precision-Recall Curves on DBLP Data (Wiki Phrases)



Precision-Recall Curves on DBLP Data (Non Wiki-phrases)

Experimental Results: Interesting Phrases Generated (From Titles & Abstracts of SIGKDD)

Query	SIGKDD	
Method	SegPhrase+	Chunking (TF-IDF & C-Value)
1	data mining	data mining
2	data set	association rule
3	association rule	knowledge discovery
4	knowledge discovery	frequent itemset
5	**time series**	decision tree
…	…	…
51	association rule mining	search space
52	rule set	domain knowledge
53	concept drift	**important problem**
54	knowledge acquisition	concurrency control
55	**gene expression data**	conceptual graph
…	…	…
201	web content	optimal solution
202	**frequent subgraph**	semantic relationship
203	intrusion detection	**effective way**
204	**categorical attribute**	space complexity
205	user preference	**small set**
…	…	…

Only in SegPhrase+

Only in Chunking

Mining Quality Phrases in Multiple Languages

- Both ToPMine and SegPhrase+ are extensible to mining quality phrases in multiple languages
- SegPhrase+ on Chinese (From Chinese Wikipedia)
- ToPMine on Arabic (From Quran (Fus7a Arabic)(no preprocessing)
  - Experimental results of Arabic phrases:

كفروا→ Those who disbelieve

بسم الله الرحمن الرحيم→ In the name of God the Gracious and Merciful

Rank	Phrase	In English
…	…	…
62	首席_执行官	CEO
63	中间_偏右	Middle-right
…	…	…
84	百度_百科	Baidu Pedia
85	热带_气旋	Tropical cyclone
86	中国科学院_院士	Fellow of Chinese Academy of Sciences
…	…	…
1001	十大_中文_金曲	Top-10 Chinese Songs
1002	全球_资讯网	Global News Website
1003	天一阁_藏_明代_科举_录_选刊	A Chinese book name
…	…	…
9934	国家_戏剧_院	National Theater
9935	谢谢_你	Thank you
…	…	…

Summary

Summary: Pattern Mining Applications: Mining Quality Phrases from Text Data

- From Frequent Pattern Mining to Phrase Mining
- Previous Phrase Mining Methods
- New Methods that Integrate Pattern Mining with Phrase Mining
  - ToPMine: Phrase Mining without Training Data
  - SegPhrase: Phrase Mining with Tiny Training Sets

Recommended Readings

- S. Bergsma, E. Pitler, D. Lin, Creating robust supervised classifiers via web-scale n-gram data, ACL'2010
- D. M. Blei and J. D. Lafferty. Visualizing Topics with Multi-Word Expressions. arXiv:0907.1013, 2009
- D.M. Blei, A. Y. Ng, M. I. Jordan, J. D. Lafferty, Latent Dirichlet allocation. JMLR 2003
- K. Church, W. Gale, P. Hanks, D. Hindle. Using Statistics in Lexical Analysis. In U. Zernik (ed.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum, 1991
- M. Danilevsky, C. Wang, N. Desai, X. Ren, J. Guo, J. Han. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents. SDM'14
- A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable Topical Phrase Mining from Text Corpora. VLDB'15
- R. V. Lindsey, W. P. Headden, III, M. J. Stipicevic. A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes. EMNLP-CoNLL'12.
- J. Liu, J. Shang, C. Wang, X. Ren, J. Han, Mining Quality Phrases from Massive Text Corpora. SIGMOD'15
- A. Parameswaran, H. Garcia-Molina, and A. Rajaraman. Towards the Web of Concepts: Extracting Concepts from Large Datasets. VLDB'10
- X. Wang, A. McCallum, X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. ICDM'07