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

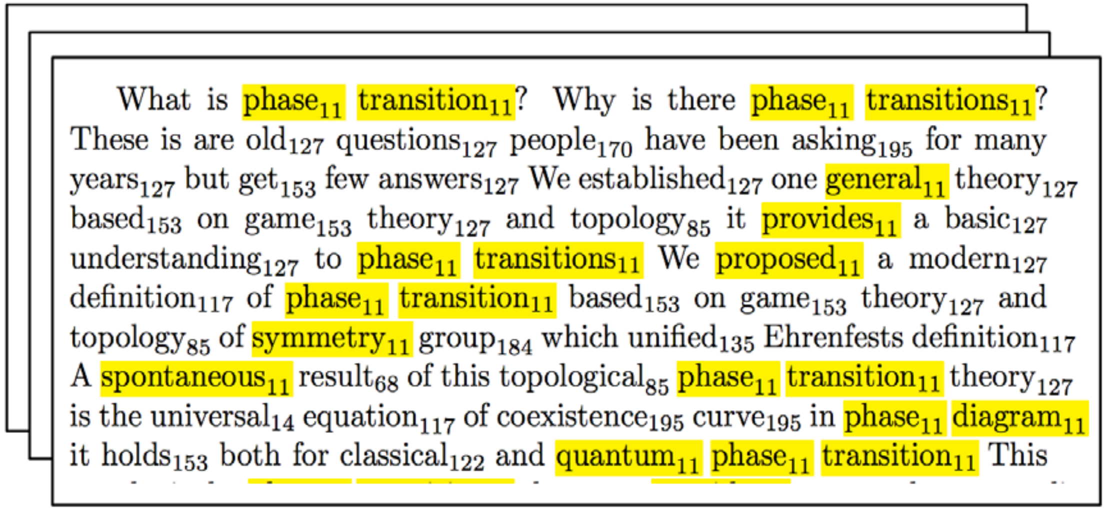
Strategy 1: Simultaneously Inferring Phrases and Topics

- 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



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