**Question 1: Phrase Mining**

1. **Topical N-grams**: Simultaneously inferring phrases and topics; Considering word order as important measure

**TurboTopics**: Post topic modeling phrase construction; Using LDA to assign topic label then merge adjacent unigrams

**KERT**: Post topic modeling phrase construction; Using frequent pattern mining and phrase ranking on popularity, discriminativeness, concordance and completeness

Why ToPMine find quality phrases and phrase-based topic: Because ToPMine first phrase construction, then topic mining. It performs frequent contiguous pattern mining to extract candidate phrases and their counts; then, performs agglomerative merging of adjacent unigrams as guided by a significance score; after than, the newly formed bag-of-phrases are passed as input to PhraseLDA that constrains all words in a phrase to each sharing the same latent topic. For phrases, it checks whether a sequence of words that occur more frequently than expected; For phrase-based topic, it uses PhraseLDA, which incorporates constraints obtained from the “bag-of-phrases” input. Because after frequent pattern mining, it would do collocation mining which comparing the observational frequencies and expected frequencies to help generating phrases. Using Phrase LDA, which incorporates constraints obtained from the “bag-of-phrases”

1. For feature extraction part, we can incorporate features using NLP technologies such as POS tagging, chunking, and semantic parsing.

Eg. Adding POS tagging, we can judge the combination of types of words. Such as form of phrase is multi nouns, verb + preposition

Eg. Using IDF to measure the semantics. “This is” has high frequent but not a phrase.

**Question 2: Entity Recognition, Typing, Embedding and Network Construction**

1. “Context-agnostic label” challenge: predict types for entity mentions regardless of contexts; select “best” label for context with specialized optimization objectives; “Type Correlation” challenge: In reality, entity types are correlated

PLE uses partial-label loss + type correlation modeling. In clean mention part, its “positive types” would be ranked higher than all “negative types”. And in noisy mention part, its “best candidate type” should be ranked higher than all “non-candidate types”. We use “best candidate type” to label that object in order to solve the problem.

1. Errors generated from subtasks, extracting typed entities and relations, may lead to more errors because wrong entity boundary makes us extract incorrect entity type and entity type error may cause relation types error. We can solve this problem by jointly extract typed entities and relations: derive entity argument types can be served as good feature for clustering relation phrases and propagate type information among entities bridges via synonymous relation phrases. They mutually enhancing each other and leads to quality recognition of unlinkable entity mentions.

First, data-driven detection of entity and relation mentions 🡪 Data-driven text segmentation and POS pattern learning from KBs; Second, Joint typing of entity and relation mentions 🡪 Noise-robust type modeling and object “translating” function to model entity-relation interactions.

Methods such as “Joint Entity and Relation Extraction using Card-Pyramid Parsing”, “Modeling Joint Entity and Relation Extraction with Table Representation” can also jointly model entity and relation types.

**Question 3: Truth Finding**

1. **Truth Finder**: Truth Finder is a HITS-like random walk method. It is a source-claim-object network framework. It computes trustworthy and confidence iteratively based on four heuristics. It uses a Source-Claim-Object network and enhance it iteratively. Limitation: quality as a single value; some sources ignore true attributes, others may produce false attributes; if there are multiple truths per entity, FN ǂ FP

**LTM**: LTM is a principled probabilistic model. It models multi-valued truth and two-sided errors. More specific, it models negative claims and two-sided source quality with Bayesian regularization. And LTM do care subtlety and demonstrates its power on truth modeling. It supports multiple true attribute values.

**GTM**: GTM is a Gaussian Truth Model for finding truth among numerical data. It mainly works on numerical data.

2. We can use LTM to solve this problem. We can treat tweets and news as different sources which tweets from certain user as a fact claim and news from certain publisher as a fact claim.

**Question 4: Spatiotemporal and Social Media Data Mining**

1. First, find a set of coarse patterns that reflect people’s semantics-level transitions; Second, split each coarse pattern into several fine-grained ones by grouping similar movement snippets

2. We can use LGTA model

First, discovery geographical topic: generate topics on regions (location of tweets) by geographic distribution of each region follows a Gaussian distribution; the words (short text message of tweets) that are close in space (location of tweets) likely belong to the same region and thus should be clustered into the same geographical topic; Second, generate a geographical document d in a collection D: sample a region (location of tweets) from the discrete distribution of region importance α; Sample location from Gaussian distribution; generate each word in document d