Semantic IR: Exploring Dependency and Word Embedding Features in Biomedical IR

Rastegar-Mojarad, M., Komandur Elayavilli, R., Liu, S., Wang, Y., Shen, F., Liu, H.

BELTracker

- Task 2: Retrieving evidence sentence for a given BEL
  - Identifying at most 10 evidence sentences
- Two architectures
  - 2015 and 2017 competitions
  - Co-occurrence based
  - IE-based
- Main components
  - Retrieval component
  - Ranking component

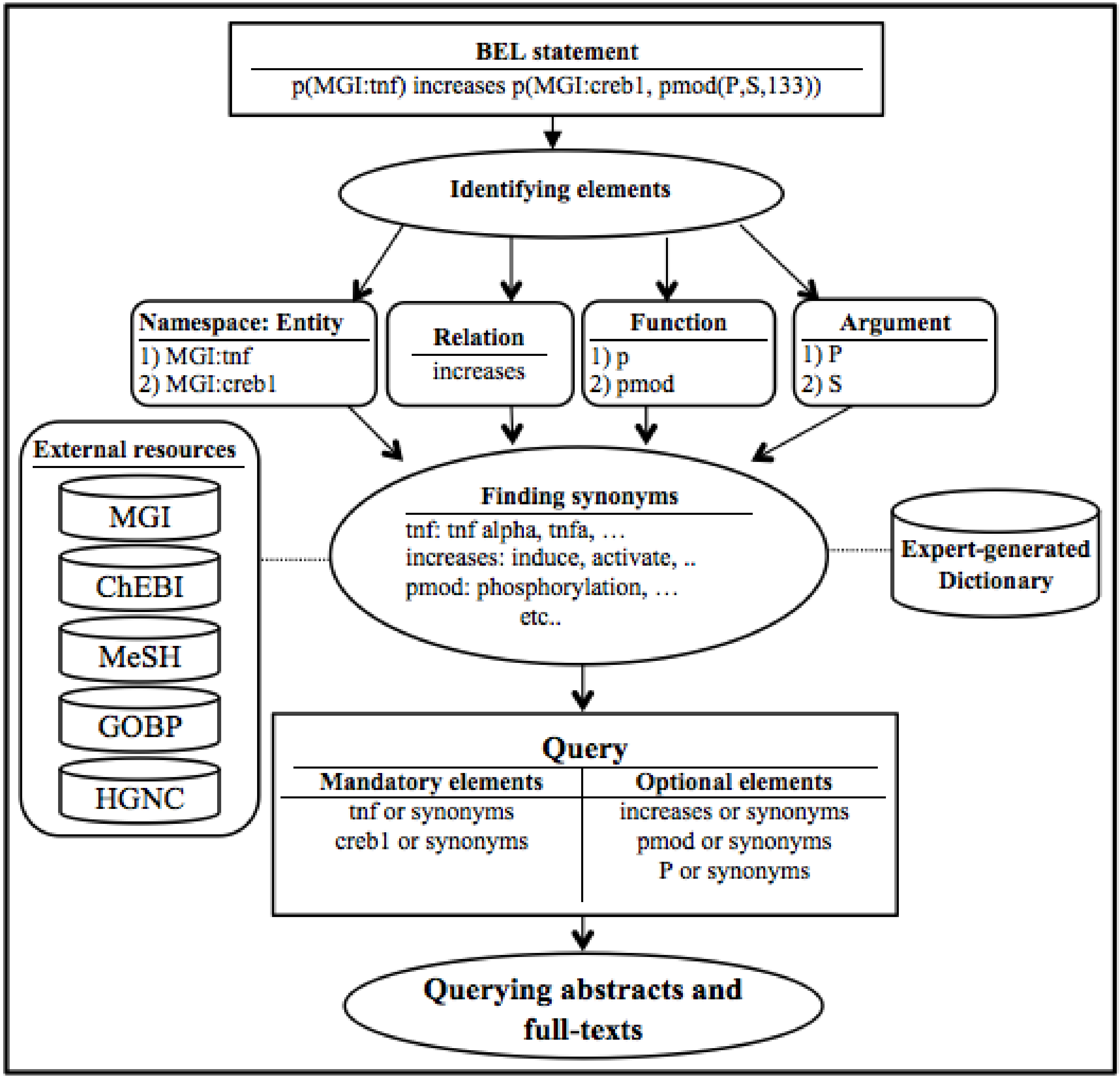
MAYO CLINIC

Co-occurrence based (2015)

- Indexing abstract and full-text
  - Elastic search

- Retrieval component
  - Translating the given BEL statement into a query
    - Identifying elements
    - Finding synonyms
      - Existing resources & expert-generated dictionaries
  - Retrieving relevant documents

MAYO CLINIC



BEL statement
p(MGI:tnf) increases p(MGI:creb1, pmod(P,S,133))
Identifying elements
Namespace: Entity
1) MGI:tnf
2) MGI:creb1
Relation
increases
Function
1) p
2) pmod
Argument
1) P
2) S
External resources
MGI
ChEBI
MeSH
GOBP
HGNC
Finding synonyms
tnf: tnf alpha, tnfa, …
increases: induce, activate, ..
pmod: phosphorylation, …
etc..
Expert-generated Dictionary
Query
Mandatory elements
tnf or synonyms
creb1 or synonyms
Optional elements
increases or synonyms
pmod or synonyms
P or synonyms
Querying abstracts and full-texts

MAYO CLINIC

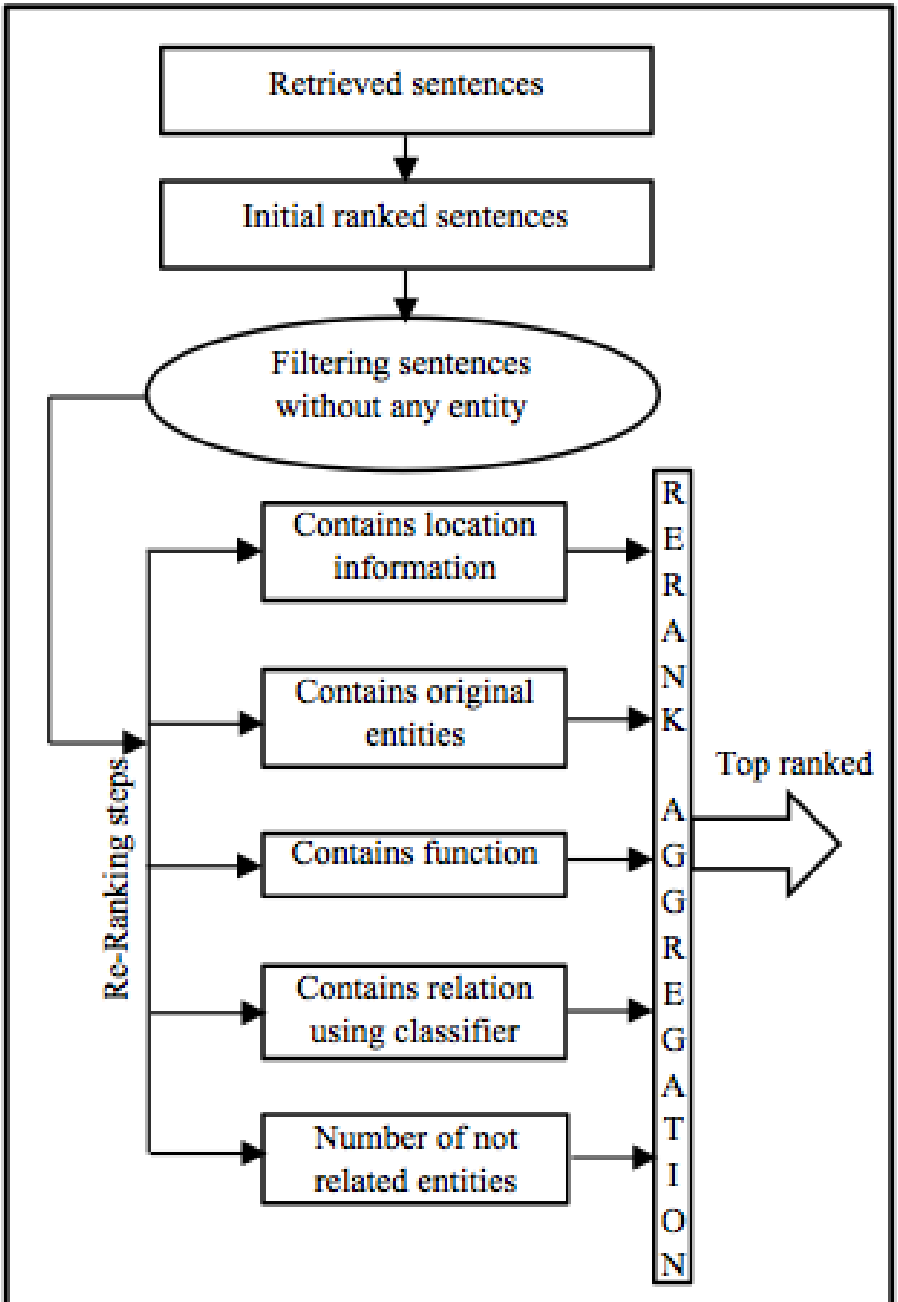
Co-occurrence based

- Ranking component

- Top 1000 results
  - Splitting into sentences
  - Removing non-relevant sentences

- Ranking sentences based on occurrence of elements
  - Manual weight assignment for elements

MAYO CLINIC



Retrieved sentences
Initial ranked sentences
Filtering sentences without any entity
Re-Ranking steps
Contains location information
Contains original entities
Contains function
Contains relation using classifier
Number of not related entities
RERANK AGGREGATION
Top ranked

MAYO CLINIC

IE-based (2017)

- Indexing only useful sentences
  - Using Semantic Medline
- Training classifiers for ranking
  - For each element
- Semantic
  - Word embedding
  - Dependency embedding

MAYO
CLINIC

Semantic Medline

- SemRep
  - Extracts semantic predications from biomedical literature
  - Biomedical entities
  - ~30 predicates
    - treat, affect, interacts with, …, **coexist with**

- Semantic Medline Database
  - A relational database
  - ~ 200 million sentences
  - ~ 90 million predications

MAYO CLINIC

Ranking component

- Using the same retrieval component
- Training classifier for ranking
  - Pair classification (BEL, sentence)
    - Positives from the training set
    - **How to define negative instances ???**
      - Randomly selecting a sentence for a BEL
        - Masking entity names
        - Not the same relation (inc. or dec.)
        - Not the same function
- Train three classifiers
  - Distant supervision

MAYO CLINIC

First classifier

- Existence of relationship
  - cat(HGNC:***SPN***) increases gtp(HGNC:***RAP1A***)
  - *RasGRP2 also catalyzes nucleotide exchange on **Rap1a** and **SPN**, but this RapGEF activity is less potent than that associated with CalDAG-GEFI.*
  - Instances (Pair of BEL statement and sentence)
    - Negative instances from Semantic medline
      - Entities: the same as randomly selected BEL
      - Predicate: **coexist with**
  - Features
    - Unigram, Bi-grams, Word-embedding

MAYO CLINIC

Second classifier

- Function

Subject	Predicate	Object
cat(HGNC:RASGRP2)	increases	gtp(HGNC:RAP1A)

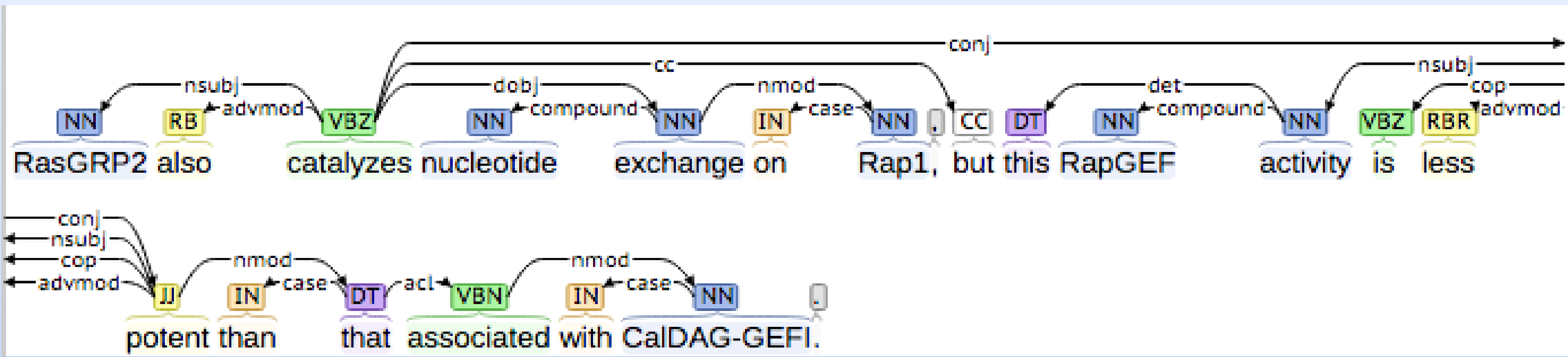
  - RasGRP2 also catalyzes nucleotide exchange on Rap1, but this RapGEF activity is less potent than that associated with CalDAG-GEFI.
  - One classifier for each function
  - Instances: BEL term+ sentence
    - Positive instances from the training data
    - Negative instances others functions

MAYO CLINIC

Second classifier

- Features
  - Surrounding terms (window of 3-5)
  - Dependency embedding

RasGRP2 also catalyzes nucleotide exchange on Rap1, but this RapGEF activity is less potent than that associated with CalDAG-GEFI.



	RasGRP2	Catalyzes
Dependency contexts	nsubj_inv_catalyzes	nsubj_RasGPR2, advmod_also, …

Third classifier & final score

- Relation
  - cat(HGNC:SPN) increases r(HGNC:CD40LG)

  - Two classes: increase and decrease
    - Instances: BEL and sentence
    - No challenge in identifying instances for each class
  - Fea. : Unigrams, bigrams, expert-generated dictionary, ..

- Final score for each retrieved sentence

$$0.40 * P_{c1} + 0.50 * P_{c2} + 0.10 * P_{c3}$$

MAYO CLINIC

Data

- Training set
  - Pair of BEL and corresponding sentence
  - 11066 BEL statements extracted from 6354 sentences
  - No entity or relation annotation

- Test set
  - 100 BEL statement for each competition
  - Manually evaluated by expert

MAYO CLINIC

Evaluation

- Precision
  - TP / (FP + TP)

- Criteria
  - ***Full***: The sentence contains the complete BEL statement.

  - ***Relaxed***: Has necessary context and/or biological background to enable extraction of BEL statement.

  - ***Context***: Even though the complete or partial BEL statement cannot be extracted from the sentence, it provides the necessary context for the BEL statement.

MAYO CLINIC

Evaluation

- Mean Average Precision (MAP)
  - For each query
    - Precisions for relevant docs considering their rank
    - Ave. the precisions : AP
  - Mean of all APs for all Qs

- Three ranking scenarios
  - Worst: All the TPs are ranked after all FPs
  - Random: Randomly reordered the results 2000 times
  - Best: All TPs are ranked before all FPs

MAYO
CLINIC

Results (2015 test data)

- Precision (co-occurrence based)

Criteria	True positive	False positive	Precision
Full	316	490	39.20
Relaxed	429	377	53.22
Context	496	310	61.53

- MAP

Criteria	Worst	Random	Co-occ.	IE-based	Best
Full	31.7	46.5	49.0	56.96	74.2
Relaxed	45.9	58.4	62.1	65.05	80.4
Context	55.2	65.7	68.9	73.15	83.5

MAYO CLINIC

Results (2017 test data)

- Two runs (top 5 evidence sentences)

Run	Criteria	True positive	False positive	Precision	MAP
ElasticSearch	Full	117	265	30.6	**59.6 (+9. 4)**
Our system	Full	121	261	**31.7 (+1.1)**	50.2
ElasticSearch	Partial	175	207	45.8	**77.5 (+0. 8)**
Our system	Partial	192	190	**50.3 (+4. 5)**	76.7

MAYO CLINIC

Discussion

- Solved issues
  - Response time
  - Extensively relied upon the lexical feature
    - Without any consideration of semantics

- Dictionaries
  - 16% of BEL entities not in the corresponding sentences
  - No results for 7 BELs
    - Focusing on Full retrieval not Partial sentences

MAYO
CLINIC

Discussion

- Only sentences in abstracts
- Ranking
  - Negative instances for the first classifier
  - No annotation for the mentions
- Evaluation
  - Need for an expert

MAYO
CLINIC

MAYO CLINIC

Thank you

Classifiers result

Classifier	Feature sets	Model	F-measure
Entity	Unigram, bi-gram, Embedding of terms between entities	SVM	0.946
Function	Window of 3-5, unigram, bigrams, Wo. Emb, Dep. Emb	SVM	0.821
Relation	Unigram, bi-grams, …	SVM	0.835

MAYO CLINIC