Generating Biological Expression Language Statements with Pipeline Approach and Different Parsers

Team Members:

Po-Ting Lai,

Ming-Siang Huang,

Wen-Lian Hsu,

Richard Tzong-Han Tsai*

Affiliation: National Central University, Taiwan

At: BioCreative VI Workshop, Bethesda, Maryland

Date: 2017/10/19

Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works

Challenges 1/2

- The task contains many stages
  - Named Entity Recognition (NER)
  - Named Entity Normalization (NEN)
  - Function Classification
  - Relation Classification.

- Developing a BEL statement generation system is more complicated than developing a single component.

Challenges 2/2

- The positions of named entity (NE), function and relation keyword are not provided in the training set.

- Therefore, the training set cannot be used to tune machine-learning models without appropriate preprocessing

BelSmile 1/4

BelSmile [1] is a pipieline BEL statement generation system, and utilizes many components including NERBio [2], NERChem[3] and RCBiosmile [4].

[1] Po-Ting Lai, Yu-Yan Lo, Ming-Siang Huang, Yu-Cheng Hsiao, Richard Tzong-Han Tsai: **BelSmile: a biomedical semantic role labeling approach for extracting biological expression language from text**. *Database* 2016 (2016)

[2] Richard Tzong-Han Tsai, Cheng-Lung Sung, Hong-Jie Dai, Hsieh-Chuan Hung, Ting-Yi Sung, Wen-Lian Hsu: **NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition.** *BMC Bioinformatics* 7(S-5) (2006)

[3] Richard Tzong-Han Tsai, Yu-Cheng Hsiao, Po-Ting Lai: **NERChem: adapting NERBio to chemical patents via full-token features and named entity feature with chemical sub-class composition**. *Database* 2016 (2016)

[4] Richard Tzong-Han Tsai, Po-Ting Lai: **A resource-saving collective approach to biomedical semantic role labeling**. *BMC Bioinformatics* 15: 160 (2014)

BelSmile 2/4

Input a sentence

"Rolipram increased phosphorylation of cAMP-response-element-binding protein (CREB) in U937 cells in a dose-dependent fashion." --- ***PMID: 10749688***

BelSmile 3/4

Step 1. entity recognition

*"**Rolipram**$_{Chemical}$ increased phosphorylation of **cAMP-response-element-binding protein**$_{protein}$ (**CREB**$_{protein}$) in U937 cells in a dose-dependent fashion."*

Step 2. entity normalization

*"**Rolipram**$_{CHEBI:rolipram}$ increased phosphorylation of **cAMP-response-element-binding protein**$_{EGID:1385}$ (**CREB**$_{EGID:1385}$) in U937 cells in a dose-dependent fashion."*

Step 3. function classification

*"**Rolipram**$_{CHEBI:rolipram}$ increased **phosphorylation**$_{p(EGID:1385,pmod(P))}$ **of cAMP-response-element-binding protein (CREB)** in U937 cells in a dose-dependent fashion."*

phosphorylation

BelSmile 4/4

Step 4. semantic role labeling

"Rolipram$_{Agent}$ *increased*$_{Predicate}$ *phosphorylation of cAMP-response-element-binding protein (CREB)*$_{Patient}$ *in U937 cells*$_{Location}$ *in a dose-dependent fashion*$_{manner}$*."*

Step 5. relation classification

"Rolipram$_{Cause}$ *increased*$_{increase}$ *phosphorylation of cAMP-response-element-binding protein (CREB)*$_{Theme}$ *in U937 cells in a dose-dependent fashion."*

Step 6. BEL statement generation

BEL statement **(Output)**

Cause	Relationship Type	Theme
a(CHEBI:rolipram)	*increases*	p(EGID:1385,pmod(P))

Our Approaches

- First, the CRFs-based gene mention recognition component was replaced by the Statistical Principle Based Approach (SPBA) NER component.
- Second, the verbal patterns were developed for function classification component.
- Third, for semantic role labeling, we ensemble our SRL parser, RCBiosmile, and a commonly-used parser, Enju.
- Lastly, our system could generate BEL statement even when the relation was presented in temporal and location statement.

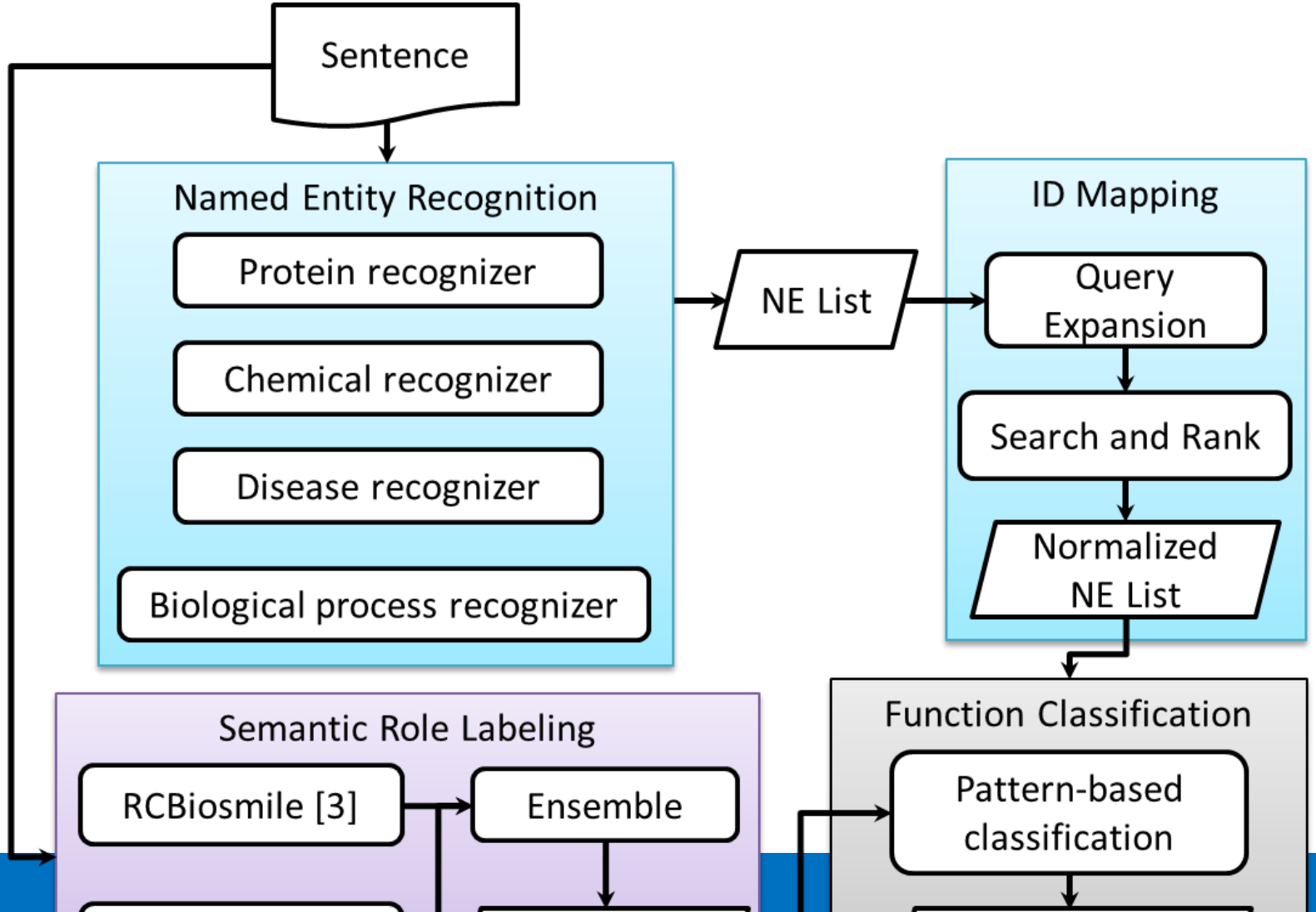
Summery of Our Approaches

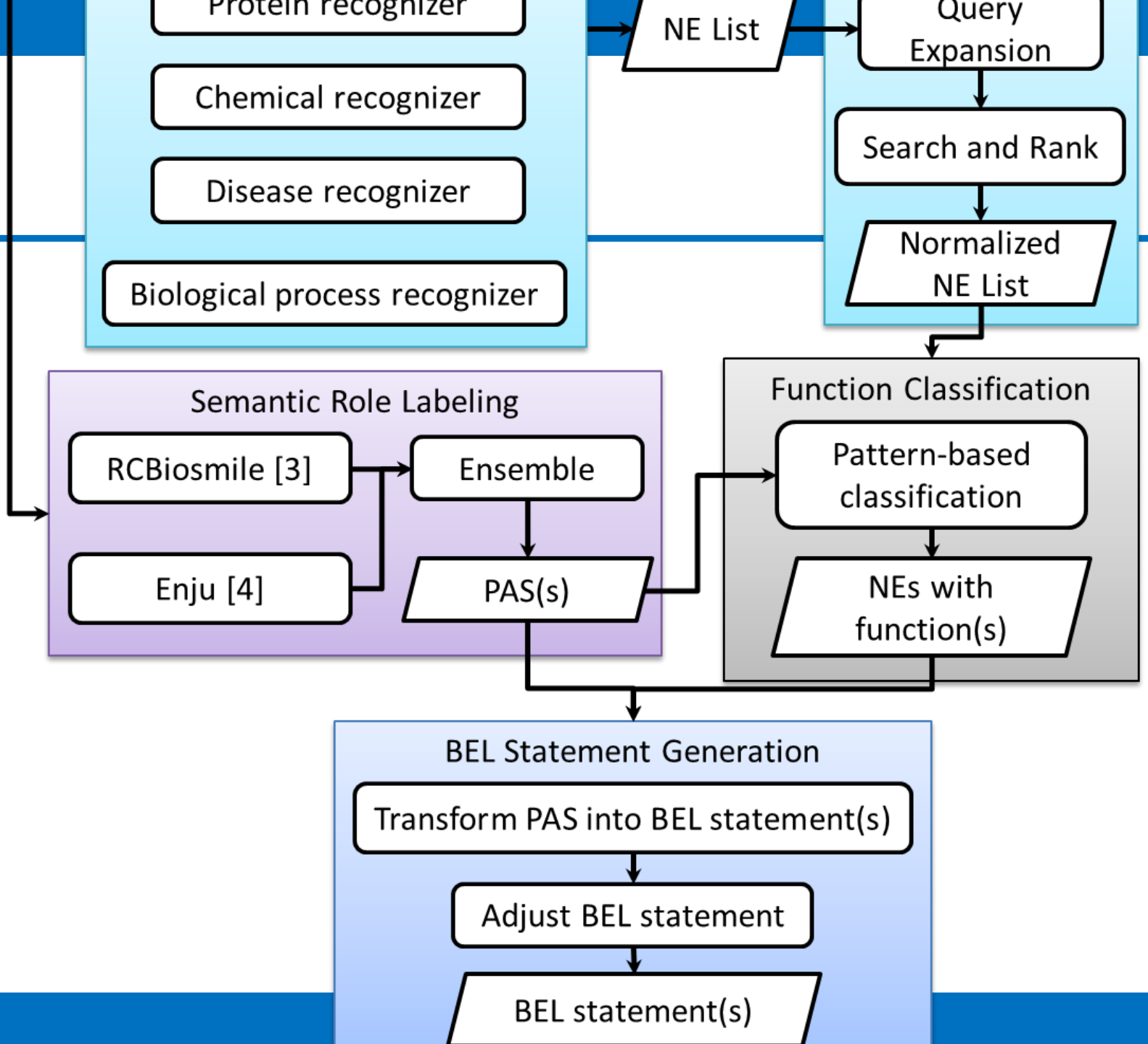
Component	Method	Training Set	Dictionary
Biological Process Recognition	Dictionary		BEL dictionary
Chemical Recognition [5]	CRFs + dictionary	BioCreative V CEMP	ChEBI
Disease Recognition	Dictionary		BEL dictionary
Protein Recognition [2]	**SPBA** + dictionary	JNLPBA + BioCreative V.5 CPRO [7]	Entrez
Function Classification	Non-verbal pattern + **verbal pattern**		
Semantic Role Labeling	RCBiosmile + **Enju**	BioProp [9]	
Relation Extraction	SRL + **time/location rules**		

Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works

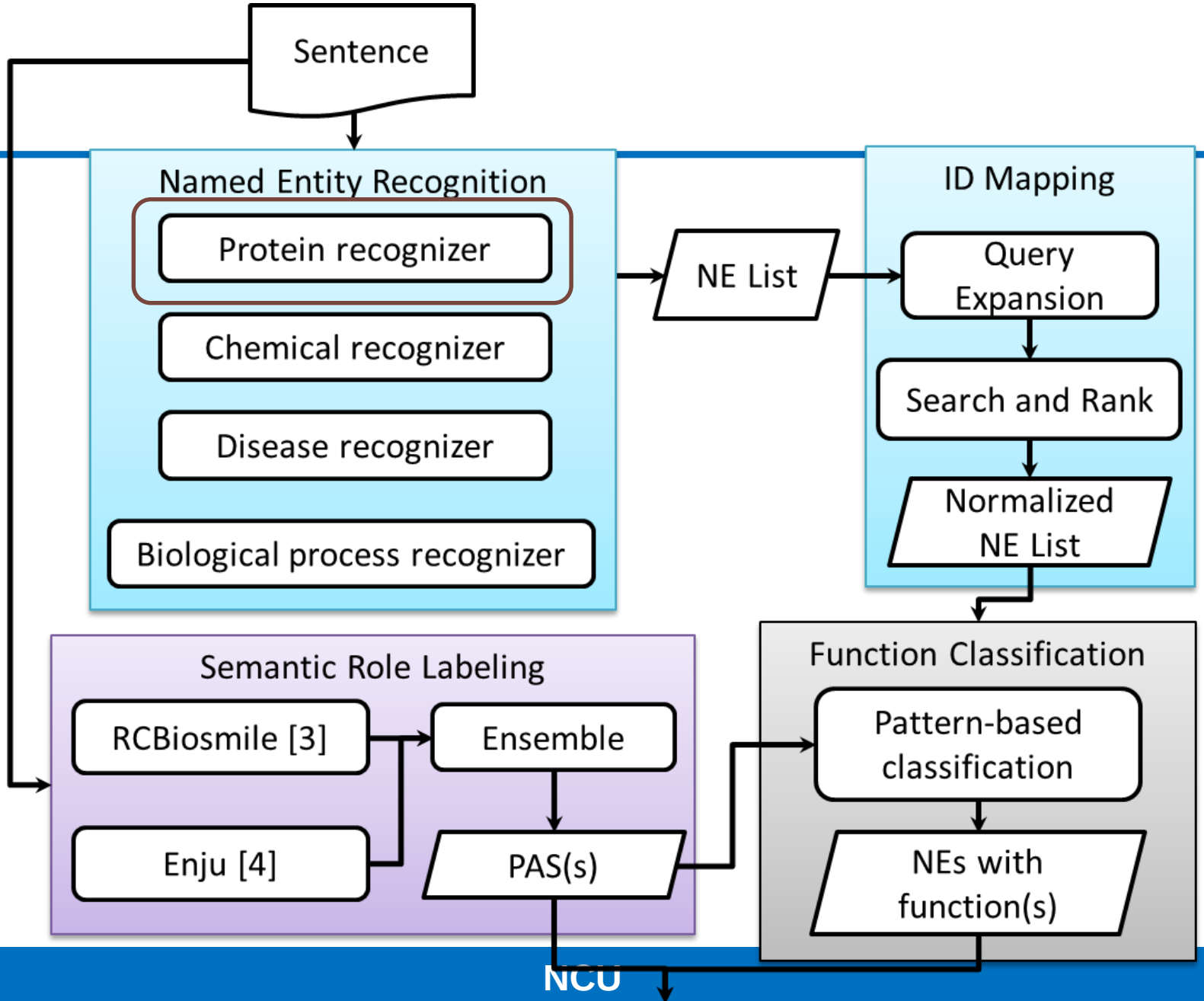
System Architecture



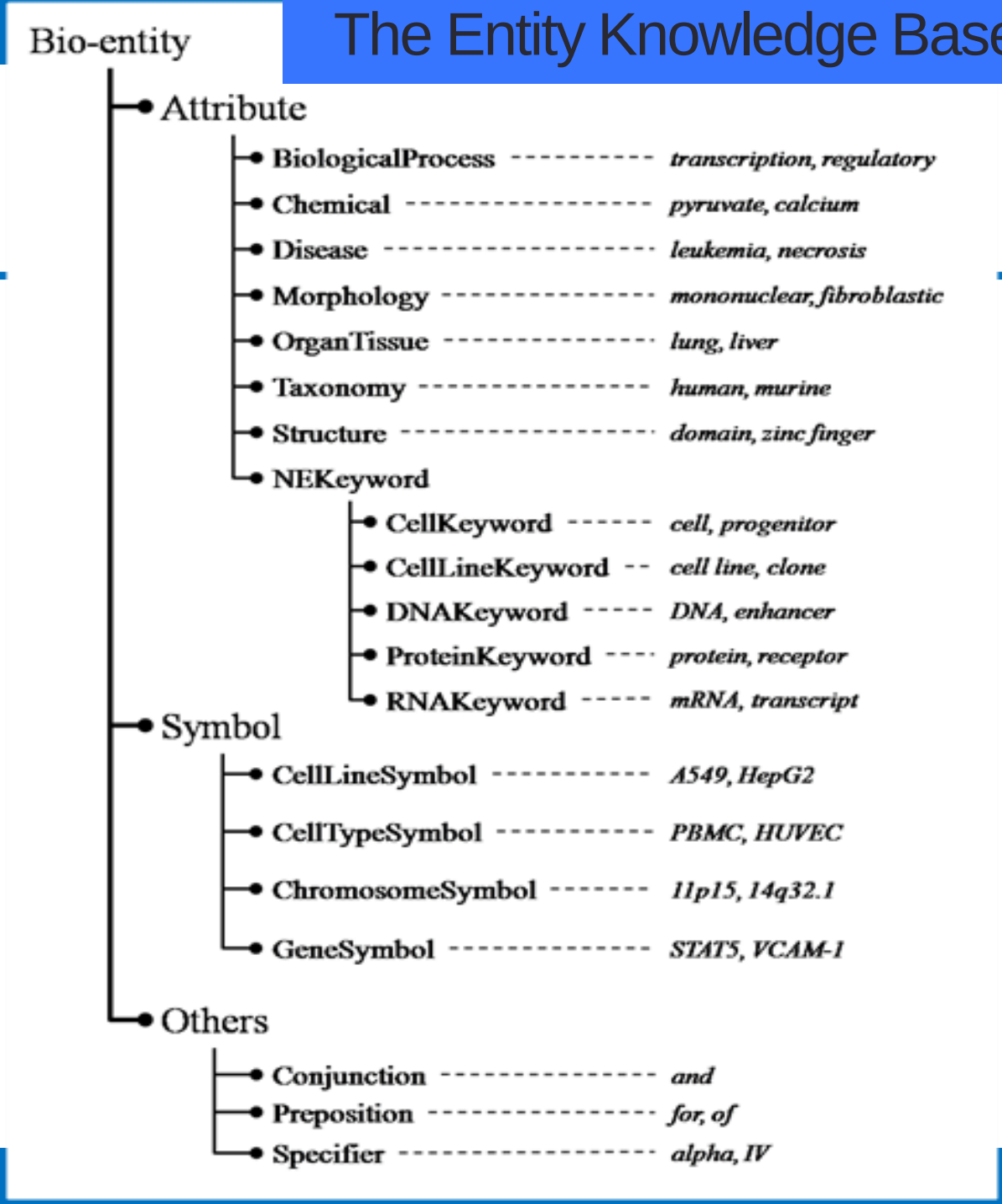

Protein recognizer
Chemical recognizer
Disease recognizer
Biological process recognizer
NE List
Query Expansion
Search and Rank
Normalized NE List
Semantic Role Labeling
RCBiosmile [3]
Enju [4]
Ensemble
PAS(s)
Function Classification
Pattern-based classification
NEs with function(s)
BEL Statement Generation
Transform PAS into BEL statement(s)
Adjust BEL statement
BEL statement(s)

Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works


Sentence
Named Entity Recognition
Protein recognizer
Chemical recognizer
Disease recognizer
Biological process recognizer
NE List
ID Mapping
Query Expansion
Search and Rank
Normalized NE List
Semantic Role Labeling
RCBiosmile [3]
Enju [4]
Ensemble
PAS(s)
Function Classification
Pattern-based classification
NEs with function(s)

The Entity Knowledge Base of SPBA

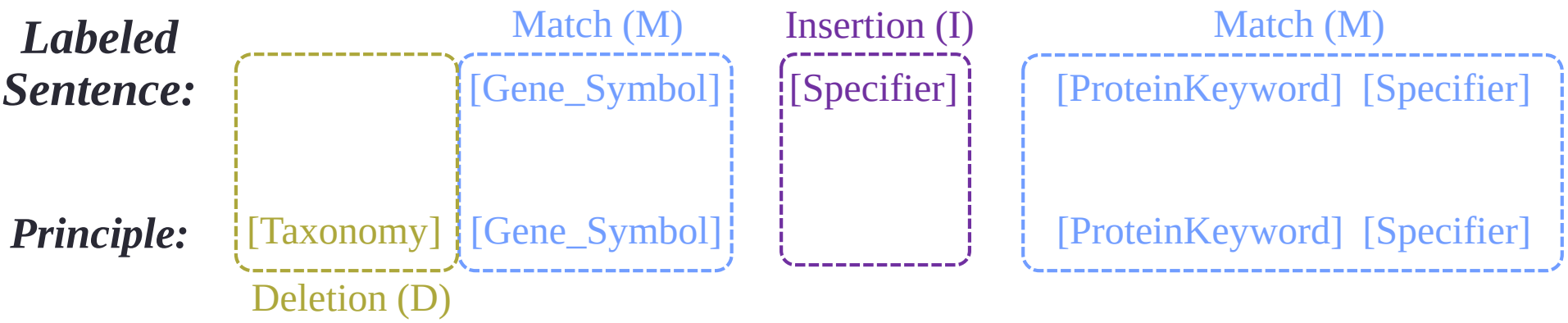


Examples of Principles

Protein Name	Principle
“***NF-kappa B***$_{GeneSymbol}$”	[GeneSymbol]
“***p50***$_{GeneSymbol}$ ***subunit***$_{ProteinKeyword}$”	[GeneSymbol][ProteinKeyword]
“***glucocorticoid***$_{Chemical}$ ***receptor***$_{FunctionKeyword}$”	[Chemical][FunctionKeyword]
“***5-lipoxygenase***$_{Enzyme}$”	[Enzyme]
“***transcription factor***$_{ProteinKeyword}$ ***NF-kappaB***$_{GeneSymbol}$”	[ProteinKeyword][GeneSymbol]

An Example of Logistic Regression-based Principle Matching

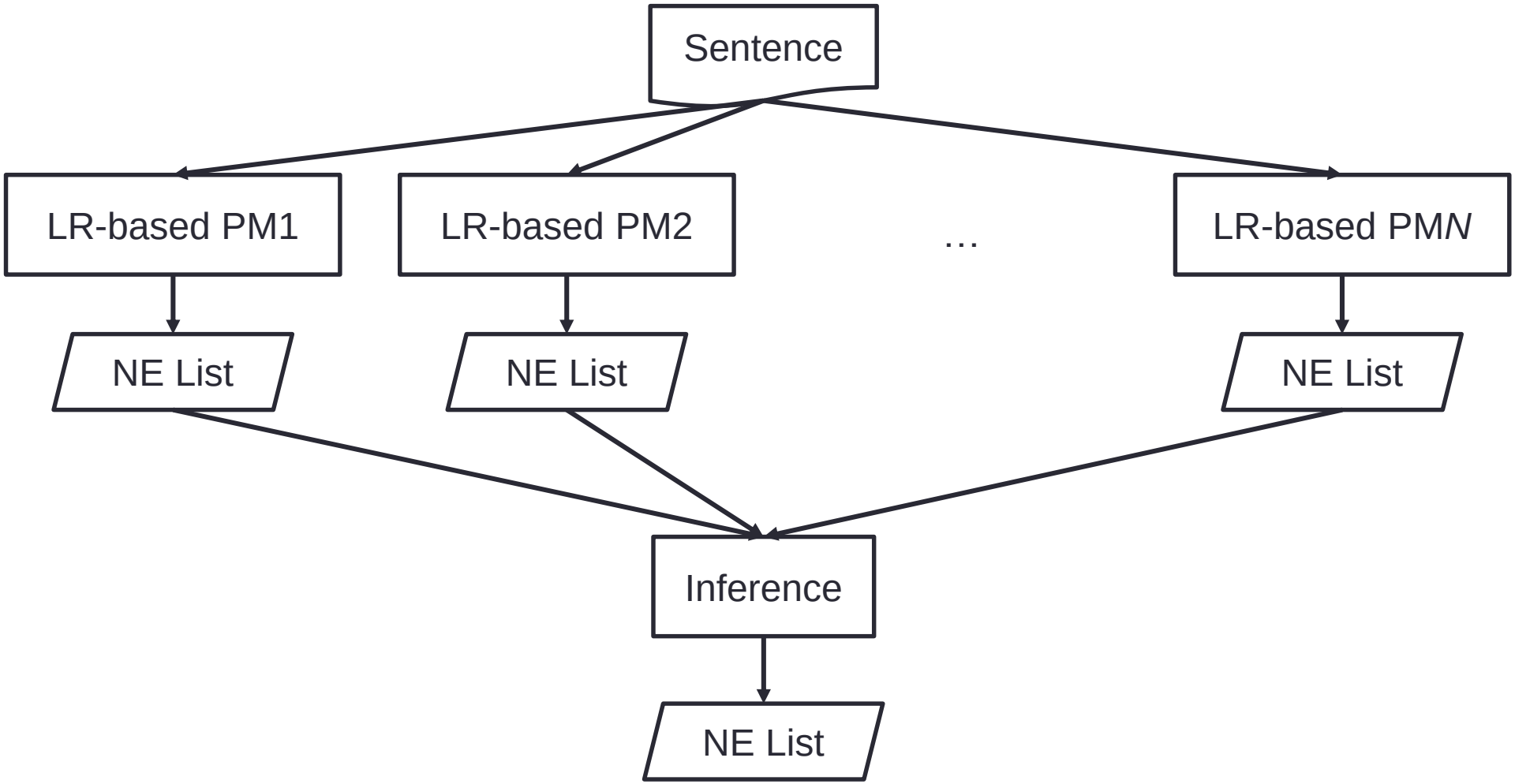
Sentence: *"… MAP3k alpha protein kinase 1…"*



Score("MAP3k alpha protein kinase 1") =

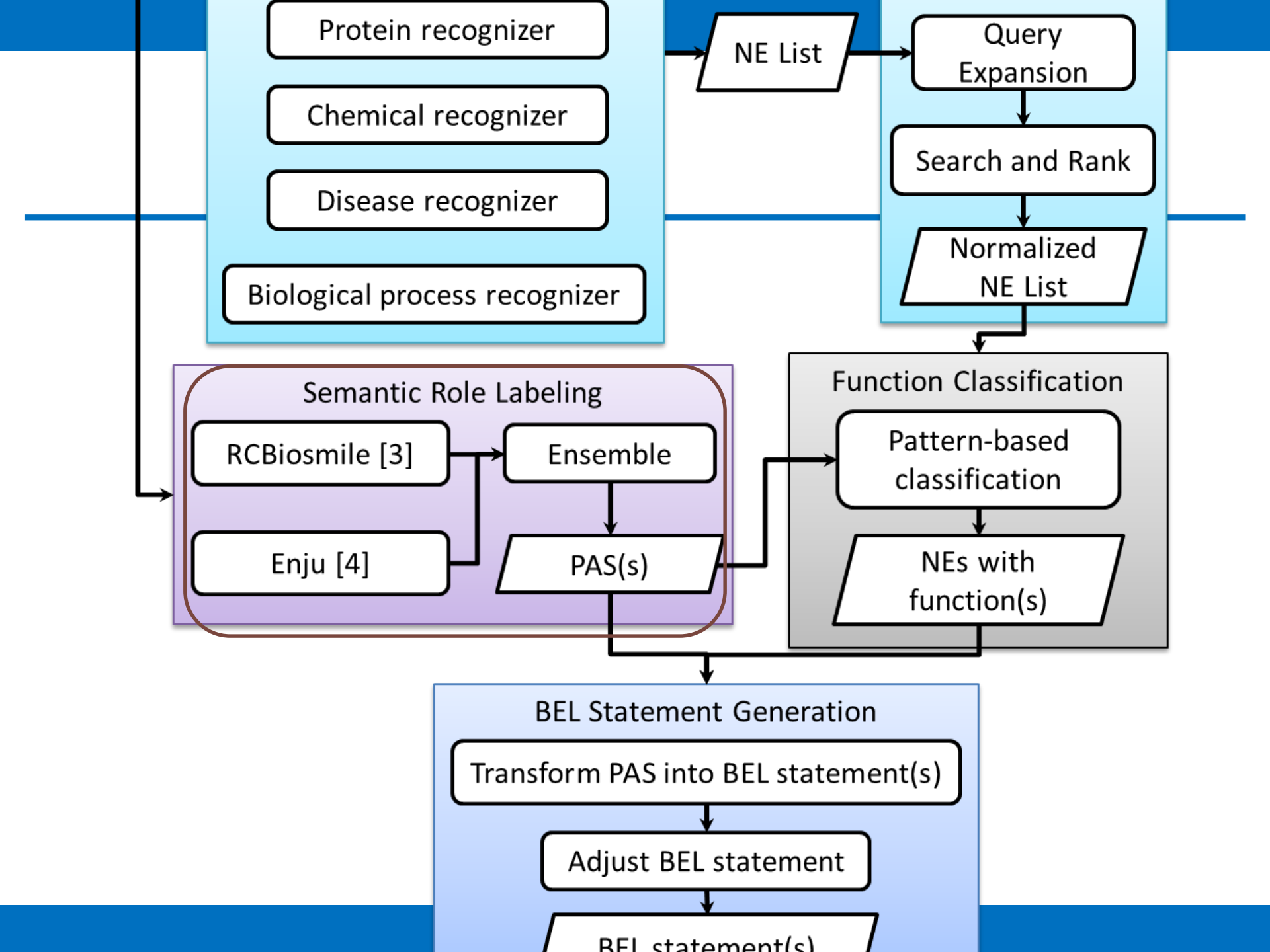
$$\lambda_D(1\_[\text{Taxonomy}]) + \lambda_M(2\_[\text{Gene\_Symbol}]) + \lambda_M([\text{Gene\_Symbol}]\_\text{"MAP3k"}) + \lambda_I([\text{Specifier}]) + \lambda_I([\text{Specifier}]\_\text{"alpha"}) + \lambda_M(3\_[\text{Protein\_End}]) + \lambda_M([\text{Protein\_End}]\text{"protein kinase"}) + \lambda_M(4\_[\text{Specifier}]) + \lambda_M([\text{Specifier}]\_\text{"1"})$$

SPBA NER


Sentence
LR-based PM1
LR-based PM2
…
LR-based PM*N*
NE List
NE List
NE List
Inference
NE List

Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works


Protein recognizer
Chemical recognizer
Disease recognizer
Biological process recognizer
NE List
Query Expansion
Search and Rank
Normalized NE List
Semantic Role Labeling
RCBiosmile [3]
Enju [4]
Ensemble
PAS(s)
Function Classification
Pattern-based classification
NEs with function(s)
BEL Statement Generation
Transform PAS into BEL statement(s)
Adjust BEL statement

Two Parsers

Input a sentence

"Rolipram increased phosphorylation of cAMP-response-element-binding protein (CREB) in U937 cells in a dose-dependent fashion." --- ***PMID: 10749688***

- **Output of RCBiosmile:**
  "Rolipram$_{Agent}$ *increased*$_{Predicate}$ *phosphorylation of cAMP-response-element-binding protein (CREB)*$_{Patient}$ *in U937 cells*$_{Location}$ *in a dose-dependent fashion*$_{manner}$*."*

- **Output of Enju:** *"Rolipram*$_{arg1}$ *increased*$_{verb\_arg12}$ *phosphorylation of cAMP-response-element-binding protein (CREB) in U937 cells in a dose-dependent fashion*$_{arg2}$*."*

Ensemble Strategy 1/2

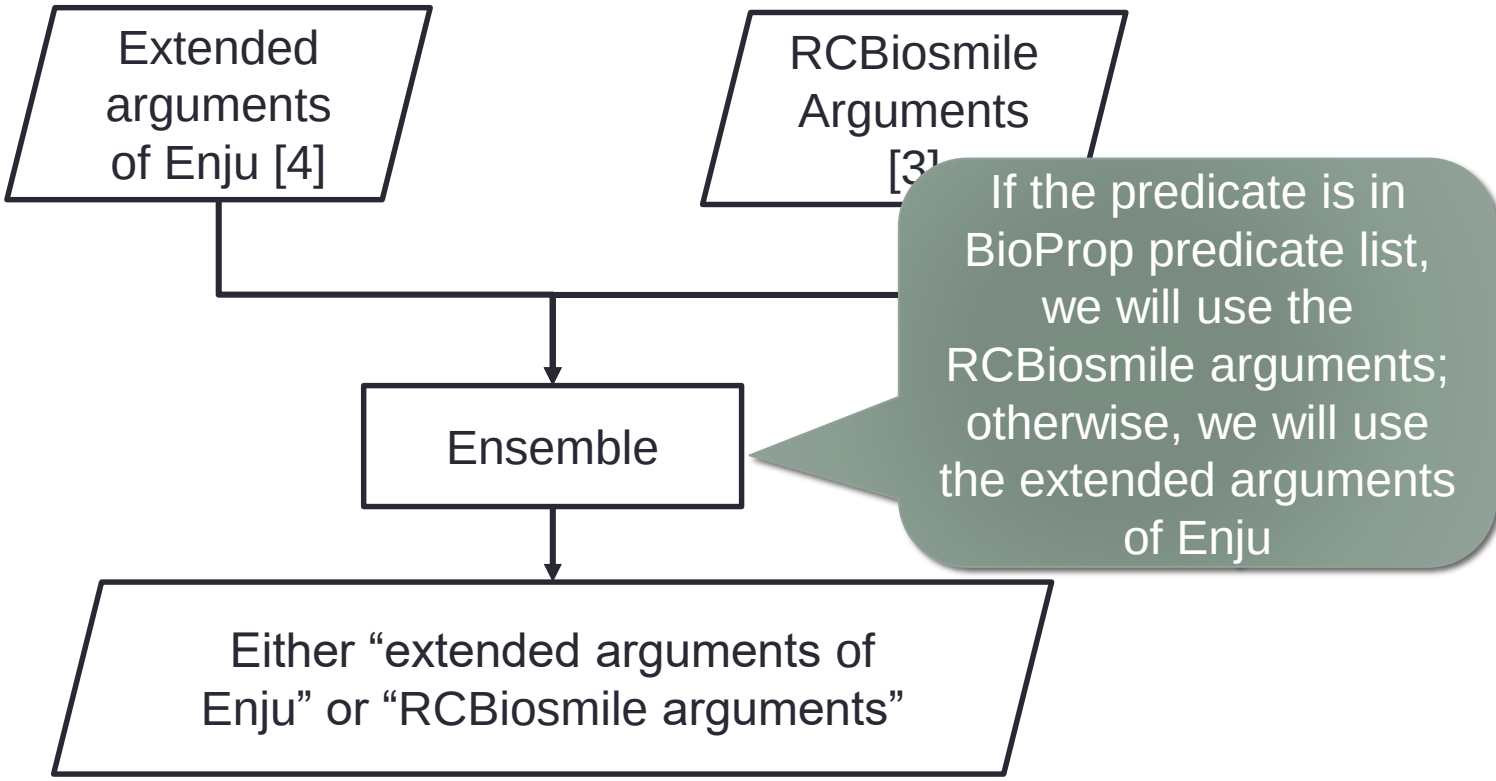
- Rule 1: Extend Enju's Arguments by RCBiosmile

Original Arguments of Enju: "*Rolipram*$_{arg1}$ *increased*$_{verb\_arg12}$ *phosphorylation of cAMP-response-element-binding protein (CREB) in U937 cells in a dose-dependent fashion*$_{arg2}$."

Extended Arguments of Enju: "*Rolipram*$_{A0}$ *increased*$_{Predicate}$ *phosphorylation of cAMP-response-element-binding protein (CREB)*$_{A1}$ *in U937 cells*$_{Location}$ *in a dose-dependent fashion*$_{manner}$"

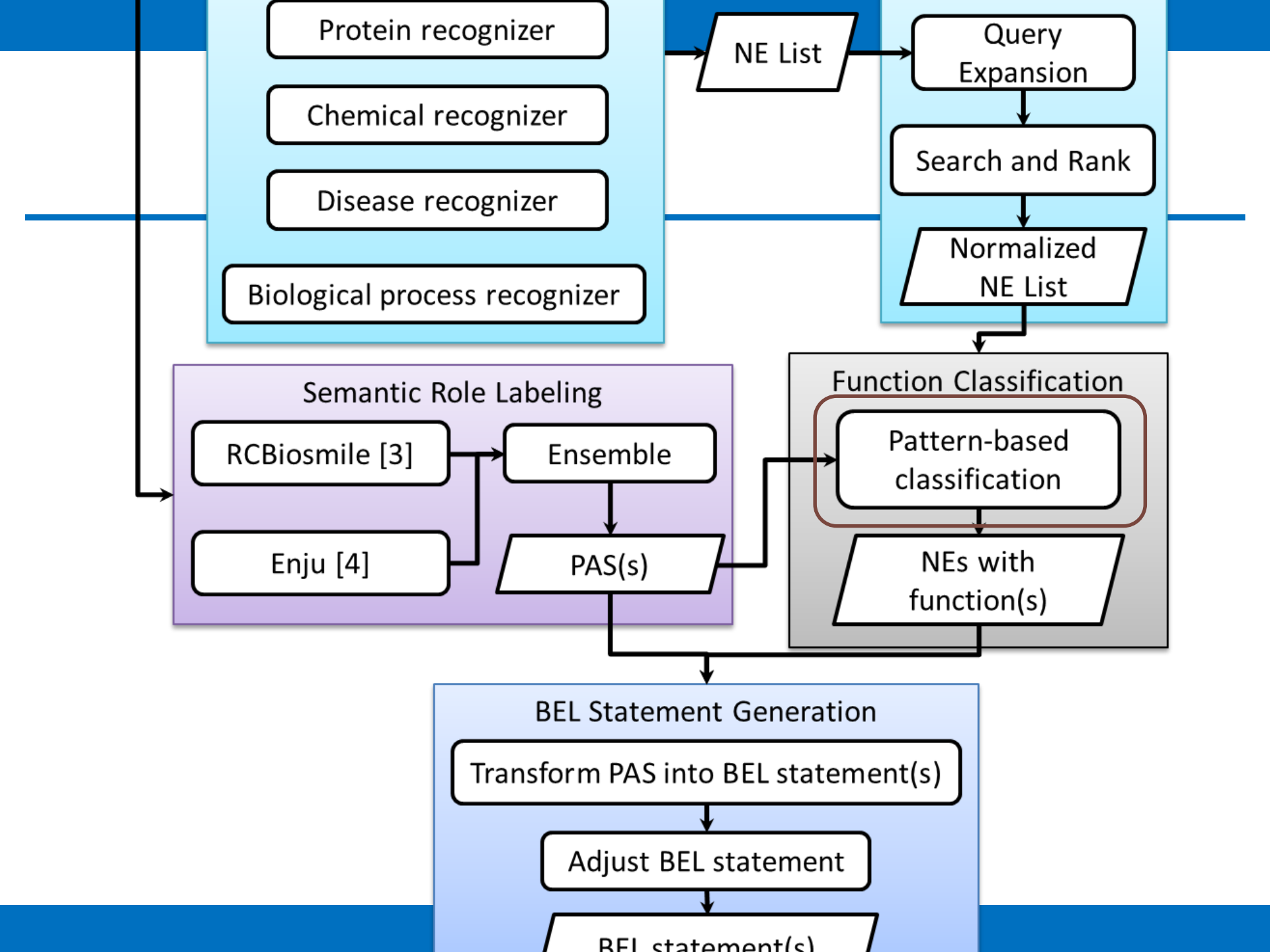
Ensemble Strategy 2/2

- Rule 2: Use predicate to decide which parser's output should be used



Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works


Protein recognizer
Chemical recognizer
Disease recognizer
Biological process recognizer
NE List
Query Expansion
Search and Rank
Normalized NE List
Semantic Role Labeling
RCBiosmile [3]
Enju [4]
Ensemble
PAS(s)
Function Classification
Pattern-based classification
NEs with function(s)
BEL Statement Generation
Transform PAS into BEL statement(s)
Adjust BEL statement

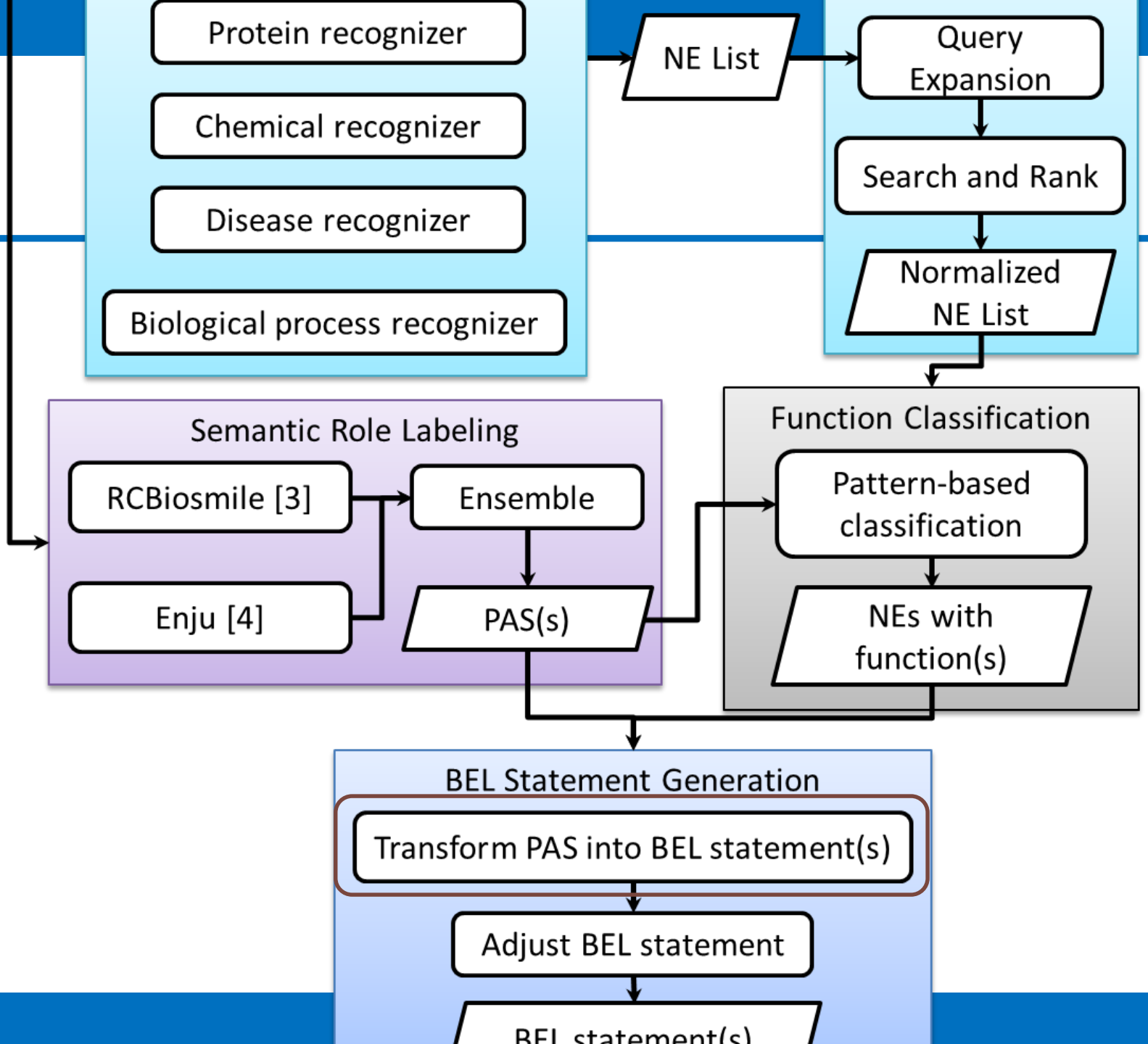
Verbal Function Pattern

- In verbal pattern, each pattern consists of predicate and arguments. We used verbal patterns to classify the functions of the NEs.
- **Example (PMID:17462626):**
  - NER: *"RalGPS2$_{EGID:55103}$ and its GEF domain activate RalA$_{EGID:5898}$ in vivo while the PH-PxxP domains inhibited it behaving as a dominant negative for the RalA pathway."*

    SRL: *"RalGPS2 and its GEF domain$_{Agent}$ activate$_{Predicate}$ RalA$_{Patient}$ in vivo while the PH-PxxP domains inhibited it behaving as a dominant negative for the RalA pathway."*
- **Verbal pattern:**
  - Agent *activate*$_{predicate}$ Patient => Cause *increase* act(Theme)
- **BEL statement:**
  - p(EGID:55103) -> act(p(EGID:5898))

Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works


Protein recognizer
Chemical recognizer
Disease recognizer
Biological process recognizer
NE List
Query Expansion
Search and Rank
Normalized NE List
Semantic Role Labeling
RCBiosmile [3]
Enju [4]
Ensemble
PAS(s)
Function Classification
Pattern-based classification
NEs with function(s)
BEL Statement Generation
Transform PAS into BEL statement(s)
Adjust BEL statement

Cause-Theme Pair Refinement

- To extract cause-theme-event relationship:
  - We map the verb predicate into either *increase* or *decrease*;
  - map the NE which is inside the agent argument into *Cause*;
  - map the NE which is inside the patient argument into *Theme*;
- However, some cause-theme-event relationships are not presented in subject-verb-object format.
  - Location statement
  - Temporal statement
- The refinement used additional rules to generate the BEL statements of them.

Location Statement

Example (PMID:20813153):

"Furthermore, the expression of Bach 2, which can form a heterodimer with mafG protein, was found to be greatly reduced, while Notch 1$_{EGID:4851}$ expression was increased in mafG$_{EGID:4097}$-deficient mice."

SRL:

"Furthermore, the expression of Bach 2, which can form a heterodimer with mafG protein, was found to be greatly reduced, while Notch 1 expression$_{Patient}$ was increased$_{Predicate}$ in mafG-deficient mice$_{Location}$."

BEL statement:

p(EGID:4097) -| p(EGID:4851)

Temporal Statement

Example (PMID:11131153):

"Furthermore the activity of lyn$_{EGID:4067}$ *kinase, evaluated by an in vitro kinase assay with enolase as a substrate, increased following IL-2*$_{EGID:3558}$ *stimulation."*

SRL:

"Furthermore the activity of lyn kinase$_{Agent}$*, evaluated by an in vitro kinase assay with enolase as a substrate, increased*$_{Predicate}$ *following IL-2 stimulation*$_{Time}$*."*

BEL statement:

act(p(EGID:3558)) -> act(p(EGID:4067))

Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works

Performances on the Stage 1

Class		Recall (%)	Precision (%)	F-score (%)
Term	Run1	50.49	84.62	**63.24**
	Run2	44.59	81.93	57.75
	Run3	46.89	88.27	61.24
Function	Run1	27.37	44.83	**33.99**
	Run2	24.21	43.4	31.08
	Run3	25.26	47.06	32.88
Relation	Run1	31.58	55.38	**40.22**
	Run2	28.07	53.33	36.78
	Run3	28.07	56.14	37.43
Statement	Run1	17.54	33.33	**22.99**
	Run2	15.35	31.82	20.71
	Run3	15.35	33.98	21.15

Run 1 used the ensemble method.
Run 2 only used the extended arguments of Enju.
Run 3 only used the RCBiosmile arguments.

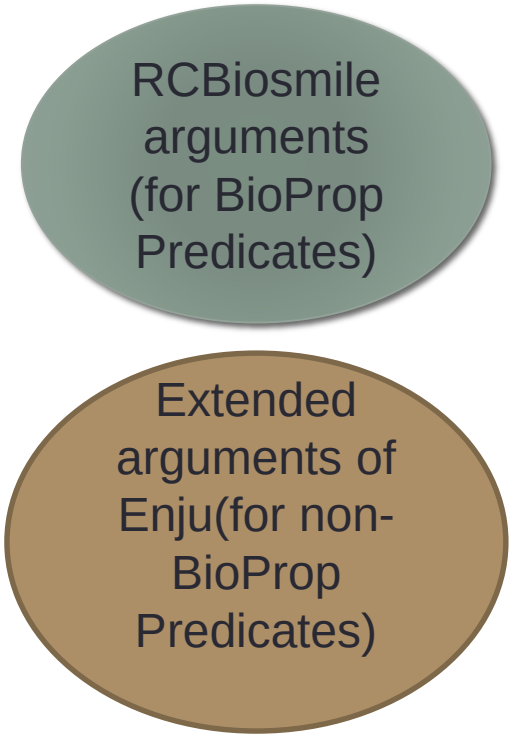
Performances on the Stage 2

Class		Recall (%)	Precision (%)	F-score (%)
Term	Run1	72.79	99.11	83.93
	Run2	76.07	99.15	**86.09**
	Run3	75.08	99.13	85.45
Function	Run1	29.47	47.46	36.36
	Run2	33.68	50.79	**40.51**
	Run3	32.63	49.21	39.24
Relation	Run1	46.93	73.29	57.22
	Run2	46.49	70.67	56.08
	Run3	47.37	73.47	**57.6**
Statement	Run1	23.68	46.15	31.3
	Run2	23.68	44.63	30.95
	Run3	24.12	46.61	**31.79**

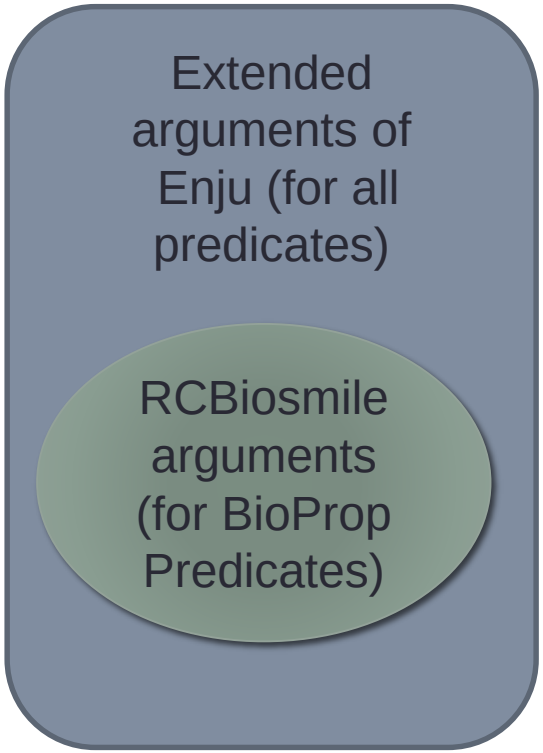
Run 1 used the ensemble method.
Run 2 used (1) all extended arguments of Enju + (2) RCBiosmile arguments (only BioProp predicates)
Run 3 used (1) all RCBiosmile augments + (2) all extended outputs of Enju (but excepts BioProp predicates)

Configurations

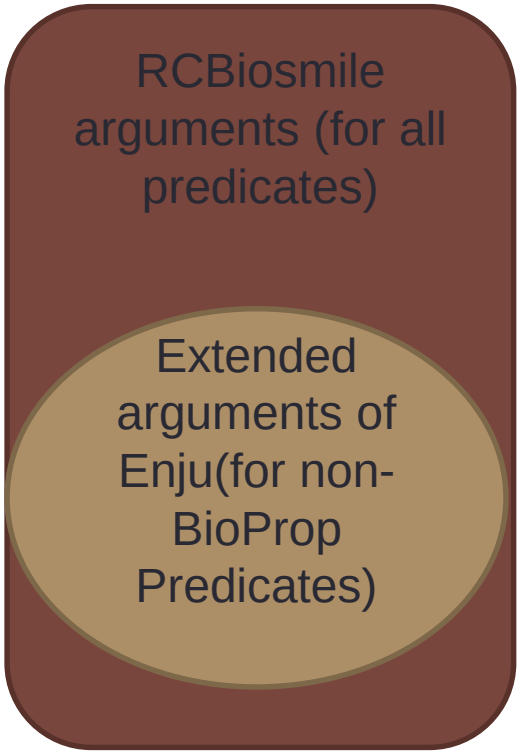
Run1



Run2



Run3 (Best)



Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works

Discussion

- **Example (PMID:7988462)**:

“Pulse-chase biosynthetic labeling studies showed that AtT-20 cells expressed much less RESP18 than the endogenous prohormone, POMC, but that glucocorticoid treatment lowered POMC and raised RESP18 biosynthetic rates so that they were nearly equimolar.”

- Two gold BEL statements should be generated.
  - “a(CHEBI:glucocorticoid) -| p(EGID:5443);”
  - “a(CHEBI:glucocorticoid) -> p(EGID:389075)”

Discussion

- **Example (PMID:7988462)**:

“Pulse-chase biosynthetic labeling studies showed that AtT-20 cells expressed much less RESP18 than the endogenous prohormone, POMC, but that glucocorticoid treatment lowered POMC and raised RESP18 biosynthetic rates so that they were nearly equimolar.”

- Two gold BEL statements should be generated.
  - “a(CHEBI:glucocorticoid) -| p(EGID:5443)”
  - “a(CHEBI:glucocorticoid) -> p(EGID:389075)”

Discussion

- **Example (PMID:7988462)**:

“Pulse-chase biosynthetic labeling studies showed that AtT-20 cells expressed much less RESP18 than the endogenous prohormone, POMC, but that glucocorticoid treatment lowered POMC and raised RESP18 biosynthetic rates so that they were nearly equimolar.”

- Two gold BEL statements should be generated.
  - “a(CHEBI:glucocorticoid) -| p(EGID:5443);”
  - “a(CHEBI:glucocorticoid) -> p(EGID:389075)”

Outline

- Challenges
- BelSmile
- Our System
  - System Architecture
  - Statistical Principle-based Approach
  - Ensemble Parsers
  - Verbal Function Pattern
  - Cause-Theme Pair Refinement
- Results
- Discussion
- Conclusion and Future Works

Conclusion and Future Works

- Using multiple components, the system performs better than using single component. Therefore, we would like to integrate different state-of-the-art systems in the future.
- In the future, we would like to apply the SPBA to tackle other NE types like chemical, disease and biological process.
- The results of the BEL statement generation are depended on individual components. We also like to use deep learning-based approaches to enhance individual components like semantic role labeling.

Thank You for Your Attention

Q & A

- **Team Members:**
- **Po-Ting Lai**,
- Ming-Siang Huang,
- Wen-Lian Hsu,
- Richard Tzong-Han Tsai

- **Affiliation:**
- National Central University, Taiwan

Thank You for Your Attention

Team Members:

Po-Ting Lai[1], Ming-Siang Huang[2], Wen-Lian Hsu[1],
Richard Tzong-Han Tsai[3*]

Affiliations:

[1] Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C.

[2] Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.

[3] Department of Computer Science and Information Engineering, National Central University, Taiwan, R.O.C.