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

Input a sentence

“Rolipram increased phosphorylation of cAMP-response-element-binding protein (CREB) in U937 cells in a dose-dependent fashion.” --- PMID: 10749688
Step 1. Entity recognition

“Rolipram\textsubscript{Chemical} increased phosphorylation of \textit{cAMP-response-element-binding protein}\textsubscript{protein} (\textit{CREB}\textsubscript{protein}) in U937 cells in a dose-dependent fashion.”

Step 2. Entity normalization

“Rolipram\textsubscript{CHEBI:rolipram} increased phosphorylation of \textit{cAMP-response-element-binding protein}\textsubscript{protein} (\textit{CREB}\textsubscript{EGID:1385}) in U937 cells in a dose-dependent fashion.”

Step 3. Function classification

“Rolipram\textsubscript{CHEBI:rolipram} increased phosphorylation of \textit{cAMP-response-element-binding protein}\textsubscript{protein} \textit{p(EGID:1385,pmod(P))} in U937 cells in a dose-dependent fashion.”
Step 4. semantic role labeling

“Rolipram\textsubscript{Agent} increased\textsubscript{Predicate} phosphorylation of cAMP-response-element-binding protein (CREB)\textsubscript{Patient} in U937 cells\textsubscript{Location} in a dose-dependent fashion\textsubscript{manner}.”

Step 5. relation classification

“Rolipram\textsubscript{Cause} increased\textsubscript{increase} phosphorylation of cAMP-response-element-binding protein (CREB)\textsubscript{Theme} in U937 cells in a dose-dependent fashion.”

Step 6. BEL statement generation

<table>
<thead>
<tr>
<th>Cause</th>
<th>Relationship Type</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>a(CHEBI:rolipram)</td>
<td>increases</td>
<td>p(EGID:1385,pmod(P))</td>
</tr>
</tbody>
</table>
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

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

- Named Entity Recognition:
  - Protein recognizer
  - Chemical recognizer
  - Disease recognizer
  - Biological process recognizer

- ID Mapping:
  - Query Expansion
  - Search and Rank
  - Normalized NE List

- Semantic Role Labeling:
  - RCBiosmile [3]
  - Ensemble

- Function Classification:
  - Pattern-based classification
• 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
The Entity Knowledge Base of SPBA

- **Bio-entity**
  - **Attribute**
    - BiologicalProcess: transcription, regulatory
    - Chemical: pyruvate, calcium
    - Disease: leukemia, necrosis
    - Morphology: mononuclear, fibroblastic
    - OrganTissue: lung, liver
    - Taxonomy: human, murine
    - Structure: domain, zinc finger
  - NEKeyword
    - CellKeyword: cell, progenitor
    - CellLineKeyword: cell line, clone
    - DNAKeyword: DNA, enhancer
    - ProteinKeyword: protein, receptor
    - RNAKeyword: mRNA, transcript
  - **Symbol**
    - CellLineSymbol: A549, HepG2
    - CellTypeSymbol: PBMC, HUVEC
    - ChromosomeSymbol: 11p15, 14q32.1
    - GeneSymbol: STAT5, VCAM-1
  - **Others**
    - Conjunction: and
    - Preposition: for, of
    - Specifier: alpha, IV
## Examples of Principles

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Principle</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;NF-kappa B_{GeneSymbol}&quot;</td>
<td>[GeneSymbol]</td>
</tr>
<tr>
<td>&quot;p50_{GeneSymbol} subunit_{ProteinKeyword}&quot;</td>
<td>[GeneSymbol][ProteinKeyword]</td>
</tr>
<tr>
<td>&quot;glucocorticoid_{Chemical} receptor_{FunctionKeyword}&quot;</td>
<td>[Chemical][FunctionKeyword]</td>
</tr>
<tr>
<td>&quot;5-lipoxygenase_{Enzyme}&quot;</td>
<td>[Enzyme]</td>
</tr>
<tr>
<td>&quot;transcription factor_{ProteinKeyword} NF-kappaB_{GeneSymbol}&quot;</td>
<td>[ProteinKeyword][GeneSymbol]</td>
</tr>
</tbody>
</table>
An Example of Logistic Regression-based Principle Matching

**Sentence:**
“... MAP3k alpha protein kinase 1...”

**Labeled Sentence:**
- Match (M) [Gene_Symbol]
- Insertion (I) [Specifier]
- Match (M) [ProteinKeyword] [Specifier]

**Principle:**
- Deletion (D) [Taxonomy]
- [Gene_Symbol]
- [ProteinKeyword] [Specifier]

**Score**
\[
\text{Score}("MAP3k alpha protein kinase 1") = \\
\lambda_D(1_{[\text{Taxonomy}]}) + \lambda_M(2_{[\text{Gene_Symbol}]}) + \lambda_M([\text{Gene_Symbol}]_"MAP3k") \\
+ \lambda_I([\text{Specifier}]) + \lambda_I([\text{Specifier}]_"alpha") \\
+ \lambda_M(3_{[\text{Protein_End}]}) + \lambda_M([\text{Protein_End}]_"protein kinase") \\
+ \lambda_M(4_{[\text{Specifier}]}) + \lambda_M([\text{Specifier}]_"1")
\]
SPBA NER

Sentence

LR-based PM1

LR-based PM2

LR-based PMN

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

- Output of Enju: “Rolipram$^{\text{arg1}}$ increased$^{\text{verb\_arg12}}$ phosphorylation of cAMP-response-element-binding protein (CREB) in U937 cells in a dose-dependent fashion$^{\text{arg2}}$. “
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}.”
• Rule 2: Use predicate to decide which parser’s output should be used

- **Extended arguments of Enju [4]**
- **RCBiosmile Arguments [3]**

If the predicate is in BioProp predicate list, we will use the RCBiosmile arguments; otherwise, we will use the extended arguments of Enju.
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
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: “\( \text{RalGPS2}_{\text{EGID:55103}} \) and its GEF domain activate \( \text{RalA}_{\text{EGID:5898}} \) in vivo while the PH-PxxP domains inhibited it behaving as a dominant negative for the RalA pathway.”

SRL: “\( \text{RalGPS2 and its GEF domain}_{\text{Agent}} \text{ activate}_{\text{Predicate}} \text{ RalA}_{\text{Patient}} \) in vivo while the PH-PxxP domains inhibited it behaving as a dominant negative for the RalA pathway.”

- Verbal pattern:
  - Agent \( \text{activate}_{\text{predicate}} \) Patient => Cause increase act(Theme)

- BEL statement:
  - \( p(\text{EGID:55103}) \) => \( \text{act}(p(\text{EGID:5898})) \)
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
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.
“Furthermore, the expression of Bach 2, which can form a heterodimer with mafG protein, was found to be greatly reduced, while Notch 1 expression was increased in mafG-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 was increased in mafG-deficient mice.”

BEL statement:
p(EGID:4097) \rightarrow p(EGID:4851)
Example (PMID:11131153):

“Furthermore the activity of lyn\textsubscript{EGID:4067} kinase, evaluated by an in vitro kinase assay with enolase as a substrate, increased following IL-2\textsubscript{EGID:3558} stimulation.”

SRL:

“Furthermore the activity of lyn kinase\textsubscript{Agent}, evaluated by an in vitro kinase assay with enolase as a substrate, increased following IL-2 stimulation\textsubscript{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

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

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
<th>F-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Run1</td>
<td>Run2</td>
<td>Run3</td>
</tr>
<tr>
<td>Term</td>
<td>72.79</td>
<td>76.07</td>
<td>75.08</td>
</tr>
<tr>
<td>Function</td>
<td>29.47</td>
<td>33.68</td>
<td>32.63</td>
</tr>
<tr>
<td>Relation</td>
<td>46.93</td>
<td>46.49</td>
<td>47.37</td>
</tr>
<tr>
<td>Statement</td>
<td>23.68</td>
<td>23.68</td>
<td>24.12</td>
</tr>
</tbody>
</table>

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
- RCBiosmile arguments (for BioProp Predicates)
- Extended arguments of Enju (for non-BioProp Predicates)

Run2
- Extended arguments of Enju (for all predicates)
- RCBiosmile arguments (for BioProp Predicates)

Run3 (Best)
- RCBiosmile arguments (for all predicates)
- Extended arguments of Enju (for non-BioProp Predicates)
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\textsuperscript{1}, Ming-Siang Huang\textsuperscript{2}, Wen-Lian Hsu\textsuperscript{1},
Richard Tzong-Han Tsai\textsuperscript{3*}

Affiliations:
\textsuperscript{1} Department of Computer Science, National Tsing-Hua University,
Taiwan, R.O.C.
\textsuperscript{2} Bioinformatics Program, Taiwan International Graduate Program,
Institute of Information Science, Academia Sinica,
Taipei, Taiwan, R.O.C.
\textsuperscript{3} Department of Computer Science and Information Engineering,
National Central University, Taiwan, R.O.C.