Joint Inference for Knowledge Extraction from Biomedical Literature

Hoifung Poon
Dept. Computer Science & Eng.
University of Washington

(Joint work with Lucy Vanderwende at Microsoft Research)
Outline

- Motivation
- Bio-event extraction
- Our system
- Experimental results
- Conclusion
Knowledge Extraction From Web

WWW → ...
Knowledge Extraction From Web

- **If we succeed ......**
  - Breach knowledge acquisition bottleneck
  - Semantic search, question answering, ...

- **But where should we start?**
  - More urgent and/or amenable
  - General approaches
Knowledge Extraction From Biomedical Literature

- **PubMed**: 18 million abstracts; += 2000 / mo.
- Success would mean:
  - Revolutionize biomedical research
  - Dramatic speed-up in drug design
- Grammatical English
- General challenges:
  - Beyond traditional information extraction
  - Complex, nested structures
  - Naturally call for **joint inference**
BioNLP: An Emerging Field

- Protein name recognition
- Protein-protein interaction
- Bio-event extraction: Shared task of 2009 [Kim et al. 2009]
- Pathway
- Network
BioNLP: An Emerging Field

- Protein name recognition
- Protein-protein interaction
- **Bio-event extraction: Shared task of 2009** [Kim et al. 2009]
- Pathway
- Network

This talk
This Talk: Bio-Event Extraction

Adding a few joint inference formulas to simple logistic regression doubles the F1
Outline

- Motivation
- **Bio-event extraction**
- Our system
- Experimental results
- Conclusion
Bio-Event: State change of bio-molecules

- Gene expression
- Transcription
- Protein catabolism
- Localization
- Phosphorylation
- Binding
- Regulation
- Positive regulation
- Negative regulation
Example

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...

| T1   | Protein 15 29   | p70(S6)-kinase |
| T2   | Protein 44 49   | IL-10          |
| T3   | Protein 86 90   | gp41           |

**Theme:** T1

**Cause:** T3

{_theme: Regulation, involvement: Positive_regulation}
Why Is It Hard?

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Why Is It Hard?

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1...

Traditional information extraction ignores this

*Theme*

*Cause*

*Site*

IL-10

gp41

human monocyte

Theme

activation

p70(S6)-kinase
Why Is It Hard?

Variations in denoting same events

E.g., negative regulation

532 inhibited, 252 inhibition, 218 inhibit, 207 blocked, 175 inhibits, 157 decreased, 156 reduced, 112 suppressed, 108 decrease, 86 inhibitor, 81 Inhibition, 68 inhibitors, 67 abolished, 66 suppress, 65 block, 63 prevented, 48 suppression, 47 blocks, 44 inhibiting, 42 loss, 39 impaired, 38 reduction, 32 down-regulated, 29 abrogated, 27 prevents, 27 attenuated, 26 repression, 26 decreases, 26 down-regulation, 25 diminished, 25 downregulated, 25 suppresses, 22 interfere, 21 absence, 21 repress ......
Why Is It Hard?

Same word denotes different events

E.g., appearance

“in the nucleus” $\Rightarrow$ Localization

“mRNA” $\Rightarrow$ Transcription

“IL-2 activity” $\Rightarrow$ Positive-regulation

……
## Participants

| Team         | Simple Event | Binding | Regulation | All       |
|--------------|--------------|---------|------------|-----------|
| UTurku       | 64.21 / 77.45 / 70.21 | 40.06 / 49.82 / 44.41 | 35.63 / 45.87 / 40.11 | 46.73 / 58.48 / 51.95 |
| JUlieLab     | 59.81 / 79.80 / 68.38 | 49.57 / 35.25 / 41.20 | 35.03 / 34.18 / 34.60 | 45.82 / 47.52 / 46.66 |
| ConcordU     | 49.75 / 81.44 / 61.76 | 20.46 / 40.57 / 27.20 | 27.47 / 49.89 / 35.43 | 34.98 / 61.59 / 44.62 |
| UT+DBCLS     | 55.75 / 72.74 / 63.12 | 23.05 / 48.19 / 31.19 | 26.32 / 41.81 / 32.30 | 36.90 / 55.59 / 44.35 |
| VIBGhent     | 54.48 / 79.31 / 64.59 | 38.04 / 38.60 / 38.32 | 17.36 / 31.61 / 22.41 | 33.41 / 51.55 / 40.54 |
| UTokyo       | 45.69 / 72.19 / 55.96 | 34.58 / 50.63 / 41.10 | 14.22 / 34.26 / 20.09 | 28.13 / 53.56 / 36.88 |
| UNSW         | 45.85 / 69.94 / 55.39 | 23.63 / 37.27 / 28.92 | 16.58 / 28.27 / 20.90 | 28.22 / 45.78 / 34.92 |
| UZurich      | 44.92 / 66.62 / 53.66 | 30.84 / 37.28 / 33.75 | 14.82 / 30.21 / 19.89 | 27.75 / 46.60 / 34.78 |
| ASU+HU+BU    | 45.09 / 76.80 / 56.82 | 19.88 / 44.52 / 27.49 | 05.20 / 33.46 / 09.01 | 21.62 / 62.21 / 32.09 |
| Cam          | 39.17 / 76.40 / 51.79 | 12.68 / 31.88 / 18.14 | 09.98 / 37.76 / 15.79 | 21.12 / 56.90 / 30.80 |
| UAntwerp     | 41.29 / 65.68 / 50.70 | 12.97 / 31.03 / 18.29 | 11.07 / 29.85 / 16.15 | 22.50 / 47.70 / 30.58 |
| UNIMAN       | 50.00 / 63.21 / 55.83 | 12.68 / 40.37 / 19.30 | 04.05 / 16.75 / 06.53 | 22.06 / 48.61 / 30.35 |
| SCAI         | 43.74 / 70.73 / 54.05 | 28.82 / 35.21 / 31.70 | 12.64 / 16.55 / 14.33 | 25.96 / 36.26 / 30.26 |
| U Aveiro     | 43.57 / 71.63 / 54.18 | 13.54 / 34.06 / 19.38 | 06.29 / 21.05 / 09.69 | 20.93 / 49.30 / 29.38 |
| Team 24      | 41.29 / 64.72 / 50.41 | 22.77 / 35.43 / 27.72 | 09.38 / 19.23 / 12.61 | 22.69 / 40.55 / 29.10 |
| USzeged      | 47.63 / 44.44 / 45.98 | 15.27 / 25.73 / 19.17 | 04.17 / 18.21 / 06.79 | 21.53 / 36.99 / 27.21 |
| NICTA        | 31.13 / 77.31 / 44.39 | 16.71 / 29.00 / 21.21 | 07.80 / 18.12 / 10.91 | 17.44 / 39.99 / 24.29 |
| CNB Madrid   | 50.25 / 46.59 / 48.35 | 33.14 / 20.54 / 25.36 | 12.22 / 07.99 / 09.67 | 28.63 / 20.88 / 24.15 |
| CCP-BTMG     | 28.17 / 87.63 / 42.64 | 12.68 / 40.00 / 19.26 | 03.09 / 48.11 / 05.80 | 13.45 / 71.81 / 22.66 |
| CIPS ASU     | 39.68 / 38.60 / 39.13 | 17.29 / 31.58 / 22.35 | 11.86 / 08.15 / 09.66 | 22.78 / 19.03 / 20.74 |
| U Mich       | 52.71 / 25.89 / 34.73 | 31.70 / 12.61 / 18.05 | 14.22 / 06.56 / 08.98 | 30.42 / 14.11 / 19.28 |
| PIKB         | 26.65 / 75.72 / 39.42 | 07.20 / 39.68 / 12.20 | 01.09 / 30.51 / 02.10 | 11.25 / 66.54 / 19.25 |
| Team 09      | 27.16 / 43.61 / 33.47 | 03.17 / 09.82 / 04.79 | 02.42 / 11.90 / 04.02 | 11.69 / 31.42 / 17.04 |
| KoreaU       | 20.56 / 66.39 / 31.40 | 12.97 / 50.00 / 20.59 | 00.67 / 37.93 / 01.31 | 09.40 / 61.65 / 16.31 |

Table 5: Evaluation results of Task 1 (recall / precision / f-score).
Top System: UTurku

- Adopts the pipeline architecture
- First, determines event candidates and types
- Then, classifies for each pair of candidates whether the latter is a theme or cause
- **No way to feedback information to events given evidence of arguments**
- Decisions are made independently
Joint Inference for Bio-Event Extraction

- Complex, nested structures naturally argue for joint inference
- However, under-explored for this task
- Previous best joint approach [Riedel et al. 2009] still lags UTurku by a large margin
Outline

- Motivation
- Bio-event extraction
- **Our system**
- Experimental results
- Conclusion
Design Desiderata

- Jointly predict events and arguments
- Incorporate prior knowledge, e.g.,
  - Each event has a theme
  - Only regulation events can have cause
- Expand scope of joint inference to include individual dependency edges
Markov Logic [Domingos & Lowd 2009]

- **Syntax:** Weighted first-order formulas
- **Semantics:** Feature templates for Markov nets
- A Markov Logic Network (MLN) is a set of pairs $(F_i, w_i)$ where
  - $F_i$ is a formula in first-order logic
  - $w_i$ is a real number

\[
P(x) = \frac{1}{Z} \exp \left( \sum_i w_i \cdot N_i(x) \right)
\]
Markov Logic

- Unifying framework for joint inference
- A plethora of efficient algorithms available
- Open-source implementation: Alchemy
  alchemy.cs.washington.edu
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocyte by gp41...
Joint Predictions

Involvement

Up-regulation

Prep_in

Prep_by

IL-10

gp41

Human monocyte

Activation

Prep_of

NN

p70(S6)-kinase
Joint Predictions

Involvement

- IL-10
- gp41
- Human monocyte

Active

- p70(S6)-kinase
Why Individual Dependencies?

... regulate IL-10 ...
... regulate IL-10 protein ...
... regulate IL-8 and IL-10 ...

\[ \text{regulate} \quad \text{dobj} \quad \text{IL-10} \]

\[ \text{regulate} \quad \text{dobj} \quad \text{protein} \quad \text{IL-10} \]

\[ \text{regulate} \quad \text{dobj} \quad \text{IL-8} \quad \text{IL-10} \]
Why Individual Dependencies?

Regulate IL-10... regulate IL-8 and IL-10...
Why Individual Dependencies?

... regulate IL-10 ...

... regulate IL-10 protein ...

... regulate IL-8 and IL-10 ...

regulate

dobj

IL-10

Continuation of a path ...

regulate

protein

IL-8

nn

conj

IL-10

IL-10
MLN For Bio-Event Extraction

- Logistic regression
- Hard constraints
- Linguistically motivated joint formulas
Logistic Regression

- **Lexical evidence**
  E.g.: “activation” probably refers to positive-regulation

- **Syntactic evidence**
  E.g.: “nsubj” probably leads to a cause

- **Lexical-syntactic evidence**
  E.g.: “nsubj” from “binds” probably leads to a theme
Hard Constraints

- **Events**
  E.g.: Event must have a theme

- **Argument paths**
  E.g.: If edge $s \rightarrow t$ is in a theme path, then
  either $s$ is an event
  or there is some $p \rightarrow s$ in the theme path

- **Decisions about events and argument edges**
  interdependent with each other
Linguistically-Motivated Joint Formulas

- Syntactic alternations, e.g.:
  - A increases the level of B
  - The level of B increases

- Add context-specific formula
  E.g., if *increases* signifies an event, and it has both *nsubj* and *dobj* dependencies, then *nsubj* probably leads to a cause
Correct Syntactic Error with Semantic Information

Coordination: expression of IL-8 and IL-10

expression
  └── prep_of
     └── IL-8

 conj
    └── IL-10

expression
  └── prep_of
     └── IL-8

 conj
    └── IL-10
Correct Syntactic Error with Semantic Information

**PP-attachment:** involvement of IL-8 in IL-10 regulation

```
involvement
  prep_of
  prep_in
IL-8
regulation
  nn
IL-10

involvement
  prep_of
  prep_in
IL-8
regulation
  nn
IL-10
```
Outline

- Motivation
- Bio-event extraction
- Our system
- Experimental results
- Conclusion
Dataset

- BioNLP-09 Shared Task (PubMed abstracts)
  - Training: 800
  - Development: 150
  - Test: 260

- Main evaluation criteria for the task
  - Event-level recall, precision, F1
  - Account for nested event structures
Experiment Objectives

- Relative contributions of feature components
- Identify the bottlenecks for performance
- Comparison with state-of-the-art systems
Results: Development Set

F1

LR
Results: Development Set

Add hard joint inference formulas

F1

+26

LR LR+HARD
Results: Development

Add soft joint inference formulas

|          | F1  |
|----------|-----|
| LR       | 25  |
| LR+HARD  | 55  |
| FULL     | 55 +2 |
Results: Development Set

- LR
- LR+HARD
- FULL
- NO-SYN-FIX

If no fixing syntactic errors

F1
Results: Development Set

| Method       | F1   |
|--------------|------|
| LR           | 25   |
| LR+HARD      | 35   |
| FULL         | 45   |
| NO-SYN-FIX   | 55   |
| UTurku       | 55   |
Per-Type Performance

| Event Type         | Event F1 |
|--------------------|----------|
| Catabolism         | 92       |
| Phosphorylation    | 87       |
| Expression         | 77       |
| Localization       | 75       |
| Transcription      | 71       |
| Binding            | 48       |
| Negative-Reg.      | 46       |
| Positive-Reg.      | 46       |
| Regulation         | 37       |
Per-Type Performance

| Event F1 | Trigger-Word F1 |
|----------|-----------------|
| Catabolism | 92 | 91 |
| Phosphorylation | 87 | 90 |
| Expression | 77 | 80 |
| Localization | 75 | 73 |
| Transcription | 71 | 70 |
| Binding | 48 | 71 |
| Negative-Reg. | 46 | 64 |
| Positive-Reg. | 46 | 68 |
| Regulation | 37 | 51 |
Reduce F1 error by over 10%
Compare to previous best joint approach
Future Work

- Incorporate more features
- More joint inference opportunities
- Leverage discourse (e.g., coreference)
- Joint syntactic / semantic processing
Conclusion

- First joint approach for bio-event extraction with state-of-the-art results
- Based on Markov Logic
- Novel formulation with expanded joint inference
- Correcting syntactic errors with semantic information helps