Distant supervision for relation extraction using AMR

ISI Summer 2015

with Prof Daniel Marcu & Prof Kevin Knight
Task - “reading-against-a-model”

• Big Mechanism
• BioPax (Biological Pathway Exchange) model
  • contains information about interactions between different bio entities i.e. proteins, chemicals, enzymes, small-molecules, etc.
• PubMed central
  • repository of articles that may talk about such interactions
• Task
  • Given an article, automate the process of extracting information from the article and inserting it into the model
• Index cards
  • Provides a systematic way of inserting this information
Adhesion-induced **RICTOR**-mediated **AKT** Ser473 **phosphorylation** promotes cell-survival.

- **interaction_type**: "adds_modification"
- **participant_a**: { "entity_text": "RICTOR", "entity_type": "protein", "in_model": false }
- **participant_b**: { "entity_text": "AKT", "entity_type": "protein", "features": [ { "modification_type": "phosphorylation", "site": 473 } ], "identifier": "Uniprot:AKT1_HUMAN", "in_model": true }
Relation extraction

Adhesion-induced RICTOR-mediated AKT Ser473 phosphorylation promotes cell-survival
Relation extraction

Adhesion-induced RICTOR-mediated AKT Ser473 phosphorylation promotes cell-survival
Adhesion-induced RICTOR-mediated AKT Ser473 phosphorylation promotes cell-survival
Relation extraction

Adhesion-induced **RICTOR**-mediated **AKT** Ser473 phosphorylation promotes cell-survival

Participant A

Participant B

Interaction
Relation extraction

Interaction

Participant A
[state, site..]

Phosphorylated A increased the activity of B at Ser473

Participant B
[state, site...]
Relation extraction

Participant A  Interaction  Participant B
Abstract Meaning Representation

- Semantic representation that captures “who is doing what to whom” in a sentence
- Singly rooted directed graph

The diagram represents the sentence:

```
(f / fear-01
  : ARG0 (s / soldier)
  : ARG1 (d / die-01)
  : polarity “-”)
```

The soldier was not afraid of dying
The soldier was not afraid to die
The soldier did not fear death
The association of STAT5 correlates in time with JAK1 phosphorylation and bind.
The association of STAT5 correlates in time with JAK1 phosphorylation and bind.
Example bio AMR

The association of STAT5 correlates in time with JAK1 phosphorylation and binding.
Sub-graph

The association of STAT5 correlates in time with JAK1 phosphorylation and binding.

Label: Negative
Sub-graph

The association of STAT5 correlates in time with JAK1 phosphorylation and binding.

Label: Negative
Sub-graph

The association of STAT5 correlates in time with JAK1 phosphorylation and binding.
The association of STAT5 correlates in time with JAK1 phosphorylation and binding.
The association of STAT5 correlates in time with JAK1 phosphorylation and binding.
Sub-graph

The association of STAT5 correlates in time with JAK1 phosphorylation and binding.

Label: Positive

Participant B
Baseline

- Implementation - Sahil Garg et.al.
- Generate AMR for sentence
  - Gold by humans
  - Automatic – AMR parser by Michael Pust et.al
- Generate sub-graphs containing
  - Participant A
  - Participant B
  - Interaction
- Label as positive or negative - manually
- Train a model that learns to predict a label given a sub-graph
- At test time, given a sentence generate sub-graphs and predict label
Model - Graph Kernel

- Given two sub-graphs $P_1$ and $P_2$, assign a similarity score

$$K(P_1, P_2) = 0 \text{ if } t(P_1.p, P_2.p) = 0$$

$$= k(P_1.p, P_2.p) + K_c(P_1.c, P_2.c), \text{ otherwise}$$

- $t(\cdot, \cdot, \cdot)$ is a matching function based on node type
- $k(\cdot, \cdot, \cdot)$ is similarity function using node attributes
- $K_c(\cdot, \cdot, \cdot)$ similarity function of children nodes

$$K_c(P_1.c, P_2.c) = \sum_{i,j,l(i)=l(j)} \lambda^d(i)\lambda^d(j)K(P_1[i], P_2[j]) \prod_{s=1,\ldots,l(i)} t(P_1[i_s].p, P_2[j_s].p)$$

- SVM trained on precomputed graph kernel matrices
Dataset

• Training data
  • 700 sentences (350 gold AMR and 350 auto AMR)
  • 12076 sub-graphs
  • Positive – 18%
  • Negative – 78%
  • Swap – 4%

• Test data
  • 116 sentences (gold AMR)
  • 2025 sub-graphs
  • Positive – 31%
  • Negative – 64%
  • Swap – 5%
Baseline results

![Bar Chart]

- **Baseline**
  - Precision
  - Recall
  - F1
Limitations

- Requires humans to label the training data
- Low recall of 23%
- Low percentage of positive labels in training data
Distant Supervision (DS)

- Use supervision from existing knowledge bases instead of using humans for supervision
- BioPax model contains information of the form
  - Protein A interactions with Protein B
  - Occasionally the type of interaction (but mostly increases)
  - Example
    - (DAG, PKC, increases_activity)
    - ('POU2F1', 'RUNX1', 'increases')
    - ('LEF1', 'ETS1', 'increases')
    - ('PAX4', 'JUN', 'increases')

- PubMed central contains millions of articles talking about these proteins
Distant Supervision (DS)

Extract all sentences in which the two proteins co-occur

('DAG', 'PKC')

- **DAG** is important for the activation of **PKC-β**, which phosphorylates tyrosinase, and can also be released from melanocytes through UVR action in the lipid membrane.
- It has been shown that plasma membrane lipids are also affected by UVR to release membrane-associated diacylglycerol (**DAG**), which activates **PKC-β**
- The effect of PACAP was mimicked by cAMP analogues, and abrogated by inhibitors of PKA and **PKC**, which implies that the process of caspase-3 de-activation may be mediated via both the AC/cAMP/PKA and PLC/IP3/DAG (diacylglycerol)/**PKC** signalling pathways
Filtering

- A lot of them occur a list i.e. only have nouns between them ('GATA4', 'MEIS1')
  - These genes include endoderm (CXCR4, LHX1, GATA4, GATA6), pancreas (MNX1, HNF1B, HNF1A, ONECUT1 (HNF6), PROX1, SOX9, PDX1, MEIS1, RFX6, PAX2, GLIS3, NR5A2, NKX2-2, SOX4, SOX5, KRT4, KRT18, KRT19, CPE, IAPP, SLC2A2, PLA2G1B), and numerous HOX (HOXA1, HOXA2, HOXA3, HOXA4, HOXA6, HOXB2, HOXC13, HOXD12) genes (C and A; , part B).
  - Consistent with our qPCR studies, numerous pancreas-related genes were upregulated by day 17, including TM4SF4, MAFB, CDH1 (E-CAD), SPINK1, PDX1, PROX1, GHRL, RBP4, RFX6, PCSK2, HOXA2, and HNF1A, whereas several others such as PITX2, MEIS2, DACH1, ONECUT2, JAG1, FOXA2, MEIS1, ISL1, SOX9, HNF1B, KRT19, GATA4, GATA6, and MNX1 were expressed from day 10 onward (B).
DS data

- Protein pairs from BioPax model
  - 1309
- Number of sentences extracted from PubMed articles
  - 41592
  - Only 414 (out of 1309) pairs were used
- After filtering instances in a list
  - 17446
  - Only 270 (out of 1309) pairs were used
DS label all positive

- For each sentence in DS data, auto generate AMR
- Extract sub-graphs that contain Protein A and Protein B
- Label all these sub-graphs positive
- Training data
  - Gold annotated 700 sentences
  - All positive 17446 sentences
- Test data
  - Gold annotated 116 sentences
DS all positive Result

![Graph showing Precision, Recall, and F1 scores for different sets.]
This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho.
This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho.
This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho.
This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho

**Label: Positive**
This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho.

Label: Positive
Label using AMR

This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho

Label: **Negative**
This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho.

Label: **Negative**
Label using AMR Results
Limitation

• Kernel matrix computation is too expensive for large number of sub-graphs (order of 80,000 subgraphs)
• Could not run an experiment that uses all of the 17,000 sentences that we extracted using distant supervision
• Can we build a simpler model that exploits the AMR path information?
Least Common Ancestor path in AMR

This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho.
Least Common Ancestor path in AMR

This LPA-induced rapid phosphorylation of radixin/moesin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho

Label: Positive
Feedforward neural network classifier

- For each sentence extracted using distant supervision, find the shortest path between Protein A and Protein B
- Label positive if it contains any interaction term, otherwise negative
- Generate word vectors for words in the path (word2vec model trained on PubMed central)
- Concatenate word vectors together and use it as the input layer in a feedforward neural network
- Input layer size 1600
  - word2vec vector size 100
  - Atmost 16 words in the path
  - Pad with zeros
- Hidden layer size 100
- Output layer size 1
  - > 0.5 positive
Feed forward neural network (ffNN)

Word in the AMR path via LCA
Dataset

• Training
  • Gold annotated 12076 paths from 700 sentences
  • DS data with 30,000 path from 17,000 sentences

• Test
  • Gold annotated 2025 paths from 116 sentences
Results with ffNN

- Baseline
- DS all +ve mean
- DS AMR mean
- ffNN 12K gold
- ffNN 1K DS

Precision, Recall, F1
Results with ffNN

![Graph showing results with ffNN](image)

- Baseline
- ffNN 1K DS
- ffNN 30K DS
- ffNN 12K gold + 30K DS

Key:
- Brown: Precision
- Light Blue: Recall
- Gold: F1
Future work

• Extract negative examples using distant supervision
  • Current approach has a lot more positive than negative examples
  • For pair of proteins that co-occur with high frequency but not in BioPax model, label as negative

• Extensions to ffNN
  • Add word vectors instead of concatenate
  • Softmax function in the output layer
  • Fine tune the word embedding during training

• Use the AMR edge labels
Thank you!
Questions