Distant Supervision for Cancer Pathway Extraction from Text

Presented by Gus Hahn-Powell

Hoifung Poon[1]
hoifung@microsoft.com

Kristina Toutanova[1]
Chris Quirk[1]

[1] *Microsoft Research, Redmond, WA*

October 2, 2015

How do we keep up with the literature?

- Over a million publications a year!
- Cancer pathways require a systemic understanding
- Need to bring together findings scattered across the literature

What's Distant Supervision?

- Train a classifier from a weakly labeled training set
  - This usually means noisy data (i.e. annotations that we cannot always trust)
- *supervision* comes from a knowledge base resource

Challenge

- Knowledge base is incomplete
- How to handle the noise?
- How to handle overlapping relations?

Why Distant Supervision?

- leverage existing resources (knowledge base)
- mitigate annotation sparsity

Simulated distance supervision

BioNLP 2009 Event Extraction

- Entity annotations are provided
  - training: 800 instances
  - development: 150 instances
- Only considering regulations involving proteins

Building a knowledge base from BioNLP 2009

- $R = \{$
  `positive regulation`,
  `regulation`,
  `negative regulation`,
  `NULL`
  $\}$
- Extract triples from training data sentences
  - `(Protein1:Theme, Relation, Protein2:Cause)`
  - `Relation` is conservatively labelled
    - Path to `Theme` may not have intervening `Cause`
    - When in doubt about directionality, assume `regulation`

Training the distant supervision classifier

- For each sentence in training ...
  - For each pair of proteins ...
    - Extract features and predict relation label
- For $r \in R$, r can only be assigned to a triple iff the triple exists in the database
  - *A triple's existence in the kb does not mean it must be assigned the label r (could be `NULL`)*

Features

Directionality

E_1 = theme
E_2 = cause

score	criteria
0	E_1 & E_2 overlap
1	E_1 precedes E_2
-1	E_2 precedes E_1

Distance

When E_1 follows E_2, count the distance in tokens ...

- $\texttt{if}\ (k > 5)\ 1\ \texttt{else}\ 0$
- $\texttt{if}\ (k > 10)\ 1\ \texttt{else}\ 0$
- $\texttt{if}\ (k > 15)\ 1\ \texttt{else}\ 0$
- $\texttt{if}\ (k > 20)\ 1\ \texttt{else}\ 0$

Lexical

For tokens between E_1 & E_2 ...

- Direction + words
- Direction + lemma
- Direction + each word
- Direction + each lemma

Syntactic

For the dependency path connecting E_1 & E_2 ...

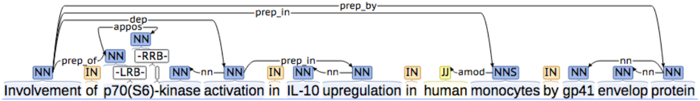

Figure: A visualization of the dependency parse for the sentence referenced in Poon et al. (2014) on page 4.

- Unlexicalized
- Lexicalized (with lemmas)
- Direction + each word
- Direction + each lemma
- path of (trigger $\rightarrow$ arg) + trigger's lemma

Distant Supervision

- Uses `MultiR` system of Hoffmann et al. (2011)
- online learning with perceptron
- 1:3 ratio for positive:negative

Choose most common label

- For all entity pairs, assign the label `positive_regulation`

Supervised system

- Some feature selection
  - filtered out features ≤ 3 occurrences in positive examples

Rules I

Data available at
`literome.azurewebsites.net/papers/psb15`

Negative Regulation

```
(ability prep_of:  (CAUSE) infmod:  (inhibit dobj:  (THEME)))
(attenuated nsubj:  (CAUSE) dobj:  (production nn:  (THEME)))
```

Positive Regulation

```
(CAUSE appos:  (factor rcmod:  (activates dobj:  (THEME))))
(CAUSE partmod:  (enhanced iobj:  (expression prep_of:  (THEME))))
```

Rules II

Skipped elements

```
walk("gene", "nn")
walk("genes", "nn")
walk("gene", "appos")
```

Comparing systems

Table 2. Test results on GENIA binary-relation classification comparing distant supervision with two baseline systems, supervised learning, and MSR11, a state-of-the-art system training on full event structures.

System	Precision	Recall	F1
Most-Frequent	3.4	69.7	6.5
Rule-Based	45.8	5.2	9.4
Distant Supervision	39.2	19.0	25.6
Supervised	37.5	29.9	33.2
MSR11	55.1	28.0	37.1

Figure: Comparing performance of different models[1].

Poon et al. (2014) doesn't attempt to capture . . .

- unary events
- "recursive events"

[1] Poon et al. (2014)

PubMed scale

- Look at relationship between cancer types and genes
- Use subset of Pathway Interaction Database (PID) to populate KB
- Extracted 1.5 million pathways
  - $800K$ were unique!
  - Estimated $372K$ are correct extractions

Challenges

- Much noisier

References I

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

Poon, H., Toutanova, K., and Quirk, C. 2014. Distant supervision for cancer pathway extraction from text. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 20, pages 120–131. World Scientific.