Seeded Discovery of Base Relations in Large Corpora

Nicholas Andrews[1] Naren Ramakrishnan[2]

[1]BBN Technologies, Cambridge, MA

[2]Virginia Tech, Blacksburg, VA

Empirical Methods in Natural Language Processing, 2008

1. **Motivation**
   - Finding connections between dissimilar documents
2. **Discovering Relations**
   - Discovering entities from seeds
   - Finding relations from co-occuring entities
   - Identifying base relations
3. **Experiments**
   - PPI sentence identification
   - Comparison with supervised methods
   - Base relation identification
4. **Discussion**

Finding connections between unrelated documents

Motivation

- **Problem**: given two seemingly unrelated concepts, find connections between them
- Building a *story* between them, "storytelling"

Building stories

An algorithm for storytelling at the document level

- Step 1: Build a document graph $G = (V, E)$ where vertices V are documents and edges exists between each pair of documents $v_1, v_2 \in V$ iff $sim(v_1, v_2) > \alpha$ for some threshold α.
- Step 2: Search (e.g., A^*) starting at the start documents
- Step 3: Rank stories according to some measure of "connectivity"

Building stories

Searching at the document level

- The good: only need a measure of similarity between documents
- The bad:
  - no guarantee of connections at the entity and relationship level
  - difficult to summarize results!

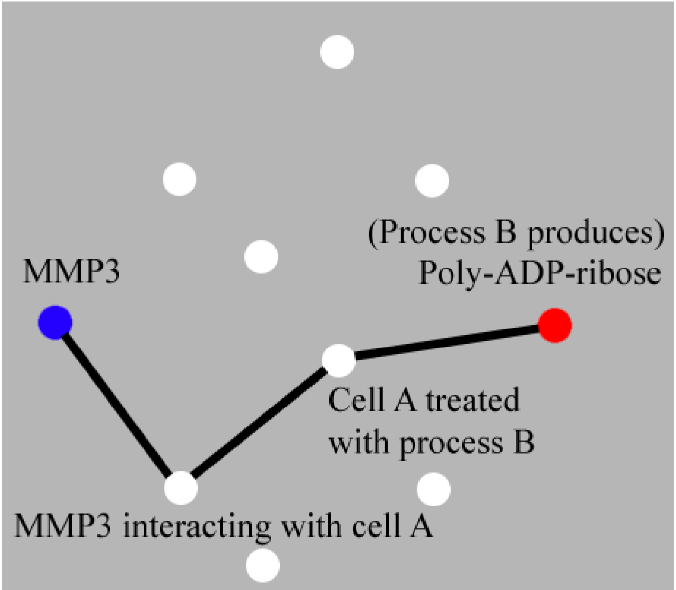
From document level to sentence level

Goal

- Model stories at the sentence level instead of the document level: make a graph where vertices are entities and edges represent relations between them . . .
- . . . but do so with minimal supervision: i.e., no PoS tagging, no parsing, no NER

How far can you get at the sentence level without any supervision?

A biomedical concept graph



Relationship discovery vs. relationship extraction

Relationship discovery: what is an edge?

- Input: Entities
- Output: Relations

Relationship extraction: build the entire concept graph

- Input: Relations, entities
- Output: More relations and entities

Relationship discovery

Method overview

- Expand an initial set of seed entities
- Identify pairs of entities likely to be in some relation
- Group relations together

Frequency patterns for entity extraction

Expanding seed entities

- Frequency meta-patterns: symbol H matches any high frequency word, symbol L matches any low frequency word (Davidov, 2006)
- Assumption: frequent words are unlikely to be content words

Example

LHL matches “apples and oranges” but not “not my apples”

Using frequency patterns to expand seeds

Example

"apples and oranges"

Building a set of fruits F

- We know that apples are fruits: start with a set $F = (apples)$
- Encounter "apples and oranges": recognize "apples"
- If we understand *and*, then it is a good indicator that oranges $\in F$!

Properties of "and"

- "and" is a frequent word
- "and" is symmetric, it also works as "oranges and apples"

Finding extraction patterns

Finding extraction patterns like "and"

- Given a seed set of entities $\{E_1, E_2, ...\}$, search the corpus for phrases like $E_1 H E_2$ for any high frequency word H
- If same seeds also appear as $E_2 H E_1$, keep H as a *symmetric pattern*

Use extraction patterns to find similar entities

- Search corpus for any unfrequent word L occuring in any symmetric pattern with a seed entity, like $E_1 H L$ or $L H E_1$
- . . . then add L to set of entities
- Can be bootstrapped as more entities are added

Example extraction patterns

- HE_1HHE_2H: "for E_1 protein or E_2 protein"
- HHE_1HE_2H: "induced by E_1 or E_2 with"
- HE_1HE_2HH: "of E_1 and E_2 mrna in"

Note

We braquet the extraction pattern with high-frequency words

Accounting for noun phrases

To find relations, we look at the context between entity pairs.

Example

"melons *are larger than Granny Smith* apples"

Polluted context

- The relation is IsLarger(melons,apples), not IsLargerGrannySmith(melons,apples)
- Context is polluted with Granny Smith

Accounting for noun phrases

Chunking with frequency patterns

- Search for patterns HL^*EL^*H (where L^* stands for "zero or more of L")
- Rank chunks L^*EL^* based on the entropy of the contexts (H, H)
- **Assumption:** The more contexts a potential chunk appears in, the more "tightly" bound two words are (Shimohata, 1997)

The co-occurence assumption

From entities, find those that are in a relation.

Assumption

Frequently co-occuring entities are likely to stand in some fixed relation

Note

But if two entities occur together *n* times, it is unlikely that *all n* relation phrases express the *same* relation

Identifying relation phrases

Finding

- For each pair of entities E_1, E_2, if E_1, E_2 appear together more than β times, add each occurence to the candidate relation phrases (RPs)

Note

- Order matters! $E_1 ... E_2$ and $E_2 ... E_1$ are counted seperately

Clustering relation phrases

Why are we clustering relations?

1. To identify groups of differently expressed but semantically similar relations
2. To feed the clustering to a relation extractor to train on

The idea of a base relation

What is a base relation and why would we want to find them?

Example

induced transient increases in
induced biphasic increases in
induced an increase in
induced an increase in both
induced a further increase in

Note

Partitional clustering algorithms do not capture this property in their objective functions

Clustering relation phrases

Problem

Given candidate relation phrases R, find a subset of exemplar relations $B \subseteq R$ which optimally describe R

This is the the p-median model (PMM): given a N x N similarity matrix, find p columns such that the sum of the maximum values within each row of the selected columns are maximized

Note

The PMM can be solved optimimally for small data sets, but in general must be approximated (e.g., relaxation, VSH, **affinity propagation**)

P-median model vs partitional clustering

Comparing two algorithms.

Affinity propagation

- $O(s)$ where s is number of similarities
- does not require number of clusters as an explicit input
- Output: assignment of items to exemplars

Hierarchical agglomerative clustering

- $O(N^2 log(N))$ or $O(N^2)$ for single-linkage HAC
- does not require number of clusters as explicit input
- Output: dendogram

Experiments

Build a biomedical corpus

- Query PubMed with 25 proteins
- Keep 87300 abstracts
- 60 most frequent words considered “high frequency”, rest as potential entities

Results

Using the same 25 proteins results in:

1. about 200 symmetric extraction patterns
2. about 4500 unique single-word entities (hopefully proteins!)
3. about 3000 chunks

PPI sentence identification

Question

How well do relations identified automatically correspond with those a human would select?

Test corpus

- Biomedical abstracts marked for proteins (the entities) and protein-protein interactions (relationships)
- For each sentance in which n entities appear, build $\binom{n}{2}$ phrases

PPI sentence identification

Procedure

- Treat our identified relation phrases in aggregate.
- Mark a phrase in the test corpus positive if it includes all words of an identified relation phrase in the correct order
- Otherwise, mark it negative

Test corpora

1. **Hard corpus:** AIMED, about 1000 of 4000 are marked PPIs
2. **Easy corpus:** CB, about 2000 of 4000 are marked PPIs

2 experiments

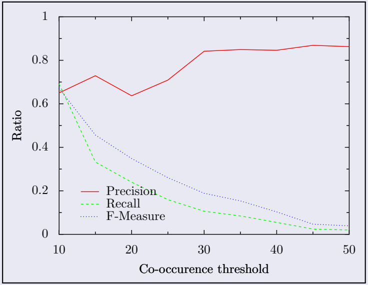
1. How are precision and recall affected by:
   1. Co-occurence threshold
   2. Minimum relation phrase length
2. How well do we do compared with supervised approaches?

Performance as entity co-occurance threshold is adjusted

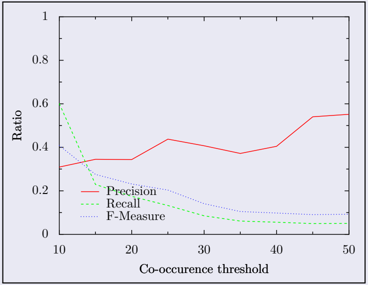
Question

Are frequently co-occuring entities more likely to be in some relationship(s)?

CB



AIMED

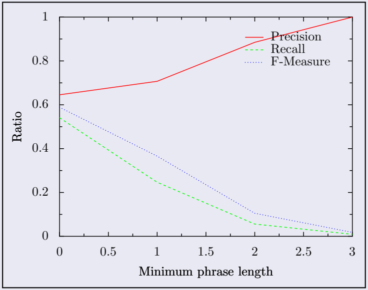


Performance as minimum RP length is adjusted

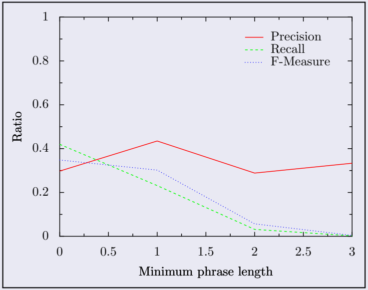
Question

How does the amount of context affect performance?

CB



AIMED



Comparison with supervised methods–AIMED corpus

At fixed parameter settings: Can we achieve the same performance as special-purpose supervised methods?

Method	P	R	F_1
RD-F_1	30.08	60.67	40.22
RD-P	**55.17**	5.04	9.25
Yakushiji et al., 2005	33.70	33.10	33.40
Mitsumori et al., 2006	54.20	42.60	47.70
Erkan et al., 2007	**59.59**	60.68	59.96

Comparison with supervised methods–CB corpus

Method	P	R	F_1
RD-F_1	65.03	69.16	67.03
RD-P	**86.27**	2.00	3.91
Erkan et al., 2007	**85.62**	84.89	85.22

Base relation identification

Question

How appropriate is the PMM for identifying base relations? (Using RD-P parameters)

Evaluation procedure by example

- Say exemplar is: **induced an increase in**
- induced transient increases in
  increases in
  induced biphasic increases in
  was induced in
  induced an increase in both
  induced biphasic *decrease* in

Base relation identification

Results

Exemplar	Size	*P* (%)
by activation of	33	87.9
was associated with	28	92.9
was induced by	24	83.3
was detected by	24	83.3
as compared with the	25	92.0
were measured with	23	87.0
mrna expression in	21	**9.5**
in response to	21	95.23
was determined by	21	90.4
with its effect in	19	**10.5**
was correlated with	18	100.0
Median precision: 86.36		

Prior work. . .

- Hasegawa *et al.*, 2004 use frequently co-occuring entities and complete-linkage HAC to identify relations in a newswire corpus (NYT 1995)
- Rosenfeld and Feldman, 2006 show that RD is an effective seed for RE
- Davidov *et al.*, 2007 use frequency patterns to extract (entity, attribute) pairs from the web

Summary

1. Frequency patterns can be used to expand seed entities and find entity chunks
2. Frequently co-occuring entities are more likely to be in some interesting relation
3. The PMM finds cluster exemplars well suited as base relations

Final notes

- Method is also applicable with seeds from multiple classes, where the goal is to find inter-class relations as well as intra-class relations

Questions

Questions?