# Relation Extraction

By. Miao Fan

**DEMO & DISCUSS**

**My proposal**

1. We can gain the metadata (common sense knowledge) from freebase to extract the relationships.
2. We can validate our algorithms on larger datasets (Wikipedia - freebase) .
3. I prefer adopting distant supervision methods (Propose new algorithm framework)

**What we need?**

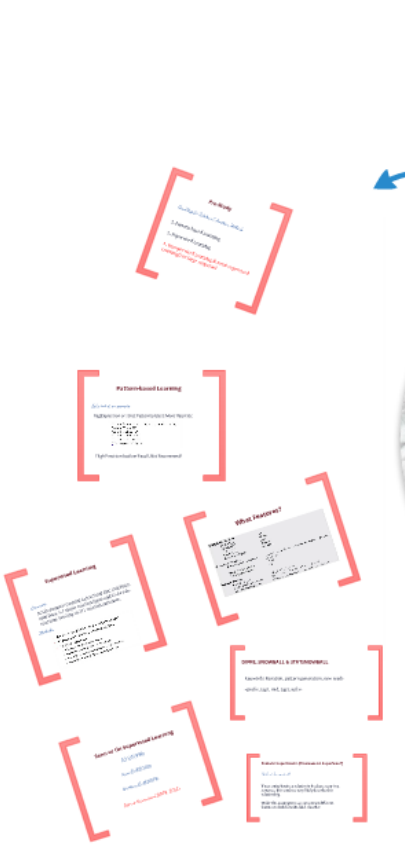
1. What kinds of relationships?
2. How do we get those relationships?
3. If the relationships are defined, what methods (Semi-supervised) do we adopt?

**What we choose?**

Pattern?

Supervised?

Semi or Unsupervised?



**Let's Begin a subtopic for digging distant supervision methods series for relation extraction**

1st article: ACL'09: Distant supervision for relation extraction without labeled data.
http://www.stanford.edu/~jurafsky/mintz.pdf

2nd article: ECML'10: Modeling relations and their mentions without labeled text.
http://marcoo.cs.umass.edu/pub/web/getpdf.php?id=911

3rd article: ACL'11: Knowledge-based weak supervision for information extraction of overlapping relations
http://homes.cs.washington.edu/~lsz/papers/hzlzw-acl2011.pdf

State-of-the-art article: ACL'12: Reducing Wrong labels in Distant Supervision for Relation Extraction.
http://www.aclweb.org/anthology-new/P/P12/P12-1076.pdf

# Relation Extraction

By. Miao Fan

**DEMO & DISCUSS**

**My proposal**

1. We can gain the metadata (common sense knowledge) from freebase to extract the relationships.
2. We can validate our algorithms on larger datasets (Wikipedia - freebase) .
3. I prefer adopting distant supervision methods (Propose new algorithm framework)

**What we need?**

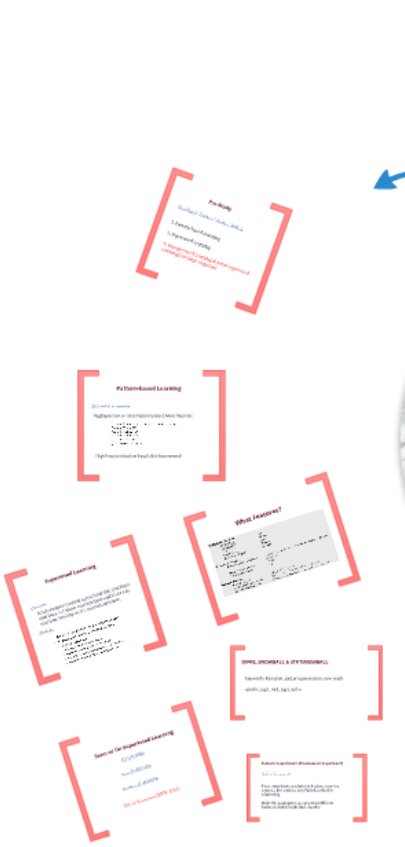
1. What kinds of relationships?
2. How do we get those relationships?
3. If the relationships are defined, what methods (Semi-supervised) do we adopt?

**What we choose?**

Pattern?

Supervised?

Semi or Unsupervised?



**Let's Begin a subtopic for digging distant supervision methods series for relation extraction**

1st article: ACL'09: Distant supervision for relation extraction without labeled data.
http://www.stanford.edu/~jurafsky/mintz.pdf

2nd article: ECML'10: Modeling relations and their mentions without labeled text.
http://marcoo.cs.umass.edu/pub/web/getpdf.php?id=911

3rd article: ACL'11: Knowledge-based weak supervision for information extraction of overlapping relations
http://homes.cs.washington.edu/~lsz/papers/hzlzw-acl2011.pdf

State-of-the-art article: ACL'12: Reducing Wrong labels in Distant Supervision for Relation Extraction.
http://www.aclweb.org/anthology-new/P/P12/P12-1076.pdf

# Pre-Study

*RoadMap for Relation Extraction Methods:*

1. Pattern-based Learning

2. Supervised Learning

3. Unsupervised Learning & Semi-supervised Learning(For large corporas)

# Pattern-based Learning

*Let's look at an example:*

RegExpression or strict Patterns(rules): More Heuristic

```
"Y such as X ((, X)* (, and|or) X)"
"such Y as X"
"X or other Y"
"X and other Y"
"Y including X"
"Y, especially X"
```

High Precision but low Recall, Not Recommend!

# Supervised Learning

*Resources:*

ACE(Automatic Content Extraction) RDC CORPORA 1000 Docs, 5-7 major relation types and 23-24 sub-relations, totaling 16.771 relation instances.

*Methods:*

- Choose a set of relations we'd like to extract
- Choose a set of relevant named entities
- Find and label data
  - Choose a representative corpus
  - Label the named entities in the corpus
  - Hand-label the relations between these entities
  - Break into training, development, and test

# What Features?

| | |
|---|---|
| **Entity-based features** | |
| $\text{Entity}_1$ type | ORG |
| $\text{Entity}_1$ head | *airlines* |
| $\text{Entity}_2$ type | PERS |
| $\text{Entity}_2$ head | *Wagner* |
| Concatenated types | ORGPERS |
| **Word-based features** | |
| Between-entity bag of words | { *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* } |
| Word(s) before $\text{Entity}_1$ | NONE |
| Word(s) after $\text{Entity}_2$ | *said* |
| **Syntactic features** | |
| Constituent path | $NP \uparrow NP \uparrow S \uparrow S \downarrow NP$ |
| Base syntactic chunk path | $NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$ |
| Typed-dependency path | $Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$ |

# Semi or Un Supervised Learning

DIPRE(1998):

SnowBall(2000):

StatSnowBall(2009):

Distant Supervision(2009, 2012):

## DIPRE, SNOWBALL & STATSNOWBALL

Keywords: Iteration, pattern generation, new seeds

<prefix, tag1, mid, tag2, sufix>

## Distant Supervision (Freebase to Supervise!!)

*Under a heuristic rule:*

If two entity having a relation in freebase occur in a sentence, this sentence very likely describe this relationship.

Under this assumption, we can extract different features to build a multi-label classifier.

**Let's Begin a subtopic for digging distant supervision methods series for relation extraction**

1st article: ACL'09: Distant supervision for relation extraction without labeled data.

http://www.stanford.edu/~jurafsky/mintz.pdf

2nd article: ECML'10: Modeling relations and their mentions without labeled text.

http://maroo.cs.umass.edu/pub/web/getpdf.php?id=931

3rd article: ACL'11: Knowledge-based weak supervision for information extraction of overlapping relations

http://homes.cs.washington.edu/~lsz/papers/hzlzw-acl2011.pdf

State-of-the-art article: ACL'12: Reducing Wrong labels in Distant Supervision for Relation Extraction.

http://www.aclweb.org/anthology-new/P/P12/P12-1076.pdf

## What we choose?

*Pattern?*

*Supervised?*

*Semi or Unsupervised?*

## What we need?

1. What kinds of relationships?

2. How do we get those relationships?

3. If the relationships are defined, what methods (Semi-supervised) do we adopt?

## My proposal

1. We can gain the metadata (common sense knowledge) from freebase to extract the relationships.

2. We can validate our algorithms on larger datasets (Wikipedia + freebase) .

3. I prefer adopting distant supervision methods (Propose new algorithm framework)

# DEMO & DISCUSS

http://www.freebase.com/m/06dfz1