

IE through Distant Supervision

Hypothesis:

If two entities maintain a certain relationship then any sentence containing these two entities is highly likely to express that ~~same~~ relationship.

Key idea:

Use a data to get lots of training examples.

Unlike → hand-crafting a few seed examples (bootstrapping)

Unlike → using hand labeled corpus (supervised setups).

Approach:

For each pair of entities obtained from a large database

- Grab sentences containing these entities from a corpus.

- Extract a lot of noisy features from these sentences.

 - Lexical, Syntactic :-

- Combine these to train a classifier.

Advantages over supervised frameworks.

- leverage existing reliable hand-crafted knowledge.

- relations have canonical names.
 { "located in", "situated in",
 "found in" }

- rich features.

Advantages on ~~types~~ unsupervised methods.

- leverage unlimited amount of data

- many weak features.

- not sensitive to the training corpus $\xrightarrow{\text{learn}}$ is domain independent.

Hypernyms for DB.

Construct a noisy training set containing occurrences from a corpus that correspond to a ~~hypernym~~ hyponym-hypersyn pair.

Shakespeare - author

from WordNet

DB

WordNet

- "... consider authors like Shakespeare..."
- "Some authors (including Shakespeare)..."
- "Shakespeare was the author of several..."
- "Shakespeare, author of The Tempest..."

Noisy examples -

- "... authors in the Shakespeare festival..."
- "The author of Shakespeare in Love is..."

Learning hyponym patterns.

① Take a corpus of sentences.
— " — doubly heavy hydrogen atom
called deuterium — — — — — "

② Collect noun pairs:
(atom, deuterium).

Count the no. of such pairs in a large
corpus (6M newswire sentences).

752311 pairs.

③ Taking pairs in each sentence
adjudge is the pair is IS-A relationship
as per the WordNet.

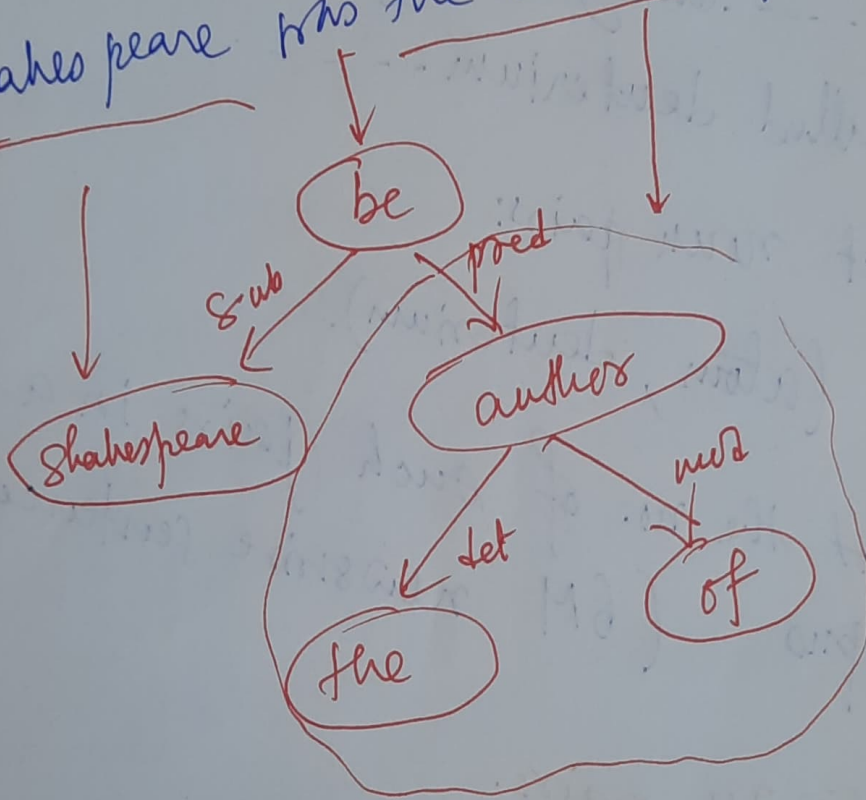
Yes / NO

④ Parse the sentences — 14,387 — extract features — 737924 — train classifier
↳ build patterns.

Syntactic dependency path between the noun pairs in the sentence.

word pair: (Shakespeare, author).

"Shakespeare was the author of several plays..."



Path:

N: sub: VBE, be, VBE: pred: N.

Noun subject
of VBE(be)

↓
Shakespeare

Noun predicate of
the VBE(be).

↓
author

N be N → NP be NP.

Stallite forms are added.

↓
such NP as NP.

- X and/or other Y
- Y such as X
- such X as Y
- Y including X
- Y, especially X
- Y like X
- Y called X
- X is Y
- X, a Y

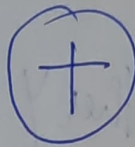
coverage??

(appositive).

Mintz, Bills, Snow, Jurafsky (2009).

Training set
Freebase.

Running corpus-
Wikipedia.



102 relations
940,000 entities
1.8 million instances.

Corpus text

Bill Gates founded Microsoft in 1975

Bill Gates, founder of Microsoft

Bill Gates attended Harvard from

Google was ~~founded~~ founded by Larry Page

Training data

(Bill Gates, Microsoft)
Label: Founder.

Feature: X founded Y.

Feature: X, founder of Y
(Larry Page, Google)

Feature: Y founded by X.

Free base:

Founder: (Bill Gates, Microsoft)

Founder: (Larry Page, Google)

College Attended: (Bill Gates, Harvard).

(Bill Gates, Harvard)
Label: College Attended.

Feature: X attended Y.

Caveat of this process?

Negative samples?

Corpus text
Larry Page took a
snipe at Microsoft.
... after Harvard invited
Larry Page to ...
Google is Bill Gates' worst fear

→ Sample 1% of unrelated pairs.

(Larry Page, Microsoft)

Label: NO-RELATION

Feature: (X took a snipe at Y.)

→ Vector.

→ lexical
→ syntactic
features.
POS patterns.
dependency
patterns.

(Larry Page, Harvard)

Label: NO-RELATION

Features: Y invited X

(Bill Gates, Google)

Label: NO-RELATION

Feature: Y is X's worst fear

(POS, dependency,
other lexical
features)

V_1	→ FOUNDER
V_2	→ COLLEGE ATTENDED
V_3	→ NO-RELATION