Recognizing Lexical Inference

Vered Shwartz

Bar-Ilan University

August 2016

Lexical Inference

- A directional semantic relation from one term to another: $x \rightarrow y$

Lexical Inference

- A directional semantic relation from one term to another: $x \to y$
- Encapsulates various relations, for example:
    - Synonymy: *(elevator, lift)*
    - Is a / hypernymy: *(pineapple, fruit), (green, color), (Obama, president)*
    - Hyponymy: *(fruit, pineapple), (color, green), (president, Obama)*
    - Meronymy: *(London, England), (hand, body)*
    - Holonymy: *(England, London), (body, hand)*
    - Causality: *(flu, fever)*

Lexical Inference

- A directional semantic relation from one term to another: $x \rightarrow y$
- Encapsulates various relations, for example:
  - Synonymy: *(elevator, lift)*
  - Is a / hypernymy: *(pineapple, fruit), (green, color), (Obama, president)*
  - Hyponymy: *(fruit, pineapple), (color, green), (president, Obama)*
  - Meronymy: *(London, England), (hand, body)*
  - Holonymy: *(England, London), (body, hand)*
  - Causality: *(flu, fever)*
- Each relation is used to infer y from x in certain contexts:
  - I ate a *pineapple* $\rightarrow$ I ate a *fruit*
  - I hate *fruit* $\rightarrow$ I hate *pineapples*
  - I visited *London* $\rightarrow$ I visited *England*
  - I left *London* $\not\rightarrow$ I left *England*

Recognizing Lexical Inference

- Given two terms, x and y, decide whether $x \rightarrow y$
  - in some senses of x and y, e.g. *apple* $\rightarrow$ *fruit*, *apple* $\rightarrow$ *company*

Example Motivation - Question Answering

Question

"What animals inhabit the Arctic regions?"

Candidate Passages

1. Polar bears inhabit the Arctic regions.
2. Indigenous people inhabit the Arctic regions.

Knowledge

bear $\rightarrow$ *animal*, but *people* $\not\rightarrow$ *animal*.

Example Motivation - Query Expansion

Query

"Actors engaged in Scientology"

Results

Tom Cruise must ditch the vile cult of Scientology NOW before
www.dailymail.co.uk/.../PIERS-MORGAN-**Tom-Cruise**-ditch-vile-cult-**Scie**...
Apr 2, 2015 - PIERS MORGAN: I wanted to be **Tom Cruise**. As a fresh-faced, 21-year-old, I watched Top Gun a dozen times at my local movie theater in ...

John Travolta Says Scientology Is A Target Because It ...
www.huffingtonpost.com/.../**john-travolta-scientology**-target_n_710268...
Apr 20, 2015 - During an interview with "Good Morning America" Monday, **John Travolta** was asked why there is so much intrigue and interest surrounding the ...

Knowledge

Tom Cruise $\rightarrow$ *actor*, *John Travolta* $\rightarrow$ *actor*.

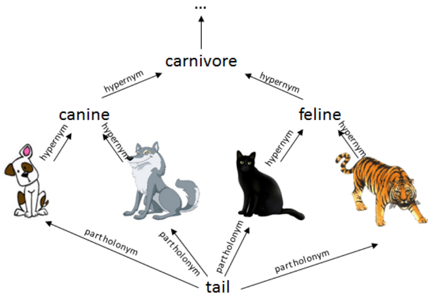
Outline

1. Resource-based Methods
   - WordNet-based Methods
   - Learning to Exploit Structured Resources for Lexical Inference
2. Corpus-based Methods
   - Distributional Approach
   - Path-based Approach
   - Integrated Path-based and Distributional Method
3. What's Next?

Resource-based Methods

Resource-based Methods for Lexical Inference

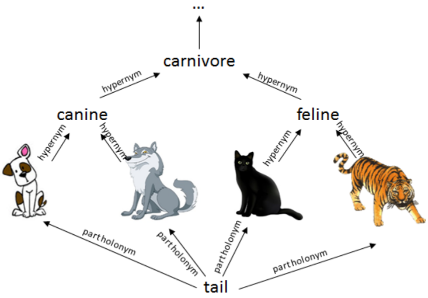
- Based on knowledge from hand-crafted resources
  - Dictionaries
  - Taxonomies (e.g. WordNet)



- Resources specify the lexical-semantic relation between terms

Resource-based Methods for Lexical Inference

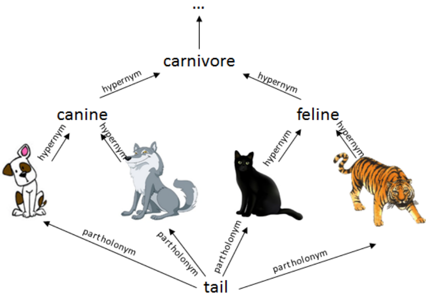
- Based on knowledge from hand-crafted resources
  - Dictionaries
  - Taxonomies (e.g. WordNet)



- Resources specify the lexical-semantic relation between terms
- The decision is based on the paths between x and y

Resource-based Methods for Lexical Inference

- Based on knowledge from hand-crafted resources
  - Dictionaries
  - Taxonomies (e.g. WordNet)



- Resources specify the lexical-semantic relation between terms
- The decision is based on the paths between x and y
- Need to predefine which relations are relevant for the task

Learning to Exploit Structured Resources for Lexical Inference

Vered Shwartz, Omer Levy, Ido Dagan and Jacob Goldberger

CoNLL 2015

WordNet-based Methods for Lexical Inference

- High precision

WordNet-based Methods for Lexical Inference

- High precision
- Limited recall:
  - WordNet is small

WordNet-based Methods for Lexical Inference

- High precision
- Limited recall:
  - WordNet is small
  - Not up-to-date
    - Recent terminology is missing: *Social Network*

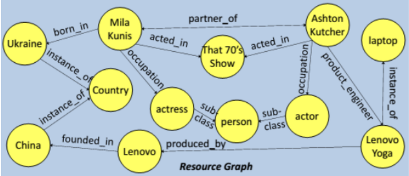
WordNet-based Methods for Lexical Inference

- High precision
- Limited recall:
  - WordNet is small
  - Not up-to-date
    - Recent terminology is missing: *Social Network*
  - Contains mostly common nouns
    - For example, it can't tell us that *Lady Gaga* is a *singer*

Community-built Resources

- Huge:
  - WIKIDATA 6M entities, 12K properties
  - DBpedia 4.5M entities, 14K properties
  - yago select knowledge 10M entities, 70 properties

  (WordNet: 150K entities, 11 properties)
- Frequently updated
- Contain proper-names



Utilizing Community-built Resources

- Idea: extend WordNet-based method using these resources

Utilizing Community-built Resources

- Idea: extend WordNet-based method using these resources
- Problem: utilizing these resources manually is infeasible
    - thousands of relations to select from!

Utilizing Community-built Resources

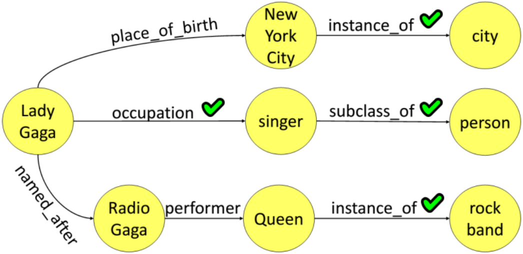
- Idea: extend WordNet-based method using these resources
- Problem: utilizing these resources manually is infeasible
  - thousands of relations to select from!
- Solution: learn to exploit these resources
  - Using genetic search

Utilizing Community-built Resources

- Input: a dataset of (x, y) term-pairs, annotated to whether $x \rightarrow y$, for a certain target lexical inference relation (e.g. "is a")

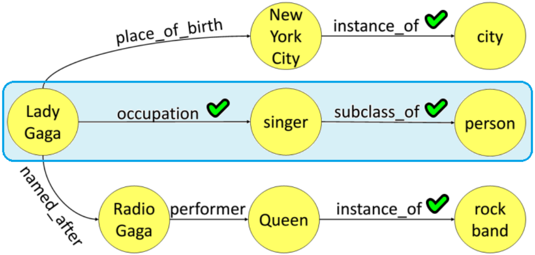
Utilizing Community-built Resources

- Input: a dataset of (x, y) term-pairs, annotated to whether $x \rightarrow y$, for a certain target lexical inference relation (e.g. "is a")
- Training: learn which properties are indicative of the target lexical inference relation



Utilizing Community-built Resources

- Input: a dataset of (x, y) term-pairs, annotated to whether $x \to y$, for a certain target lexical inference relation (e.g. "is a")
- Training: learn which properties are indicative of the target lexical inference relation



- Inference: $x \to y$ if there is a path of indicative edges from x to y

Results

- We replicate WordNet-based methods for common nouns

Results

- We replicate WordNet-based methods for common nouns
- We extract high-precision inferences including proper-names (e.g. *Lady Gaga* $\rightarrow$ *singer*)

Results

- We replicate WordNet-based methods for common nouns
- We extract high-precision inferences including proper-names (e.g. *Lady Gaga* → *singer*)
- Non-trivial resource relations are learned:

occupation	*Daniel Radcliffe* → *actor*
gender	*Louisa May Alcott* → *woman*
genre	*Touch* → *drama*
position played on sports team	*Jason Collins* → *center*

Results

- We replicate WordNet-based methods for common nouns
- We extract high-precision inferences including proper-names (e.g. *Lady Gaga* $\rightarrow$ *singer*)
- Non-trivial resource relations are learned:

occupation	*Daniel Radcliffe* $\rightarrow$ *actor*
gender	*Louisa May Alcott* $\rightarrow$ *woman*
genre	*Touch* $\rightarrow$ *drama*
position played on sports team	*Jason Collins* $\rightarrow$ *center*

- We complement corpus-based methods in high-precision scenarios, but with lower recall

Corpus-based Methods

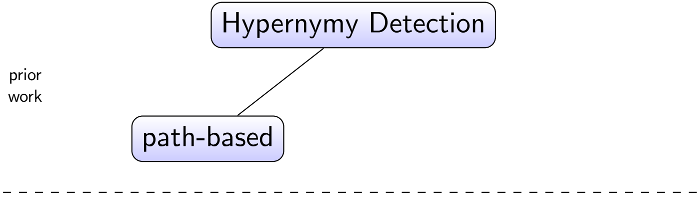
Hypernymy Detection

- Some of the following works focus on detecting hypernymy, which is a common lexical inference relation
  - the *hyponym* (x) is a type of / instance of the *hypernym* (y)
  - e.g. *(pineapple, fruit), (green, color), (Obama, president)*

Hypernymy Detection

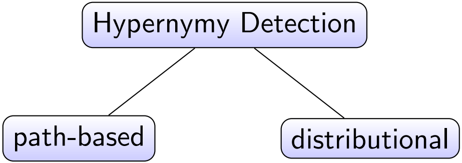
- Some of the following works focus on detecting hypernymy, which is a common lexical inference relation
  - the *hyponym* (x) is a type of / instance of the *hypernym* (y)
  - e.g. *(pineapple, fruit), (green, color), (Obama, president)*
- The hypernymy detection task: given two terms, x and y, decide whether y is a hypernym of x
  - in some senses of x and y, e.g. *(apple, fruit), (apple, company)*

Corpus-based Methods

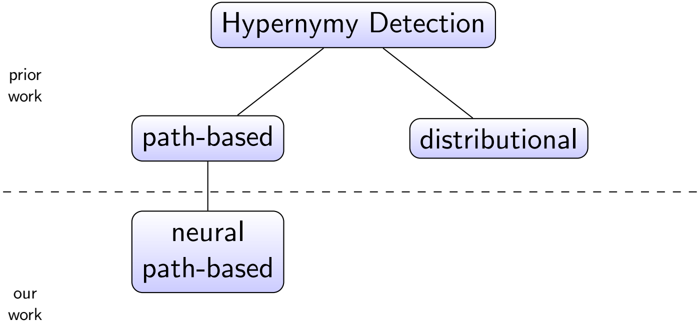


Corpus-based Methods

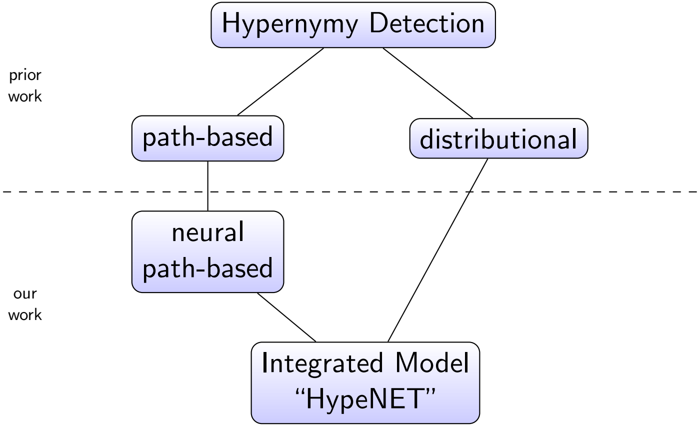
prior
work



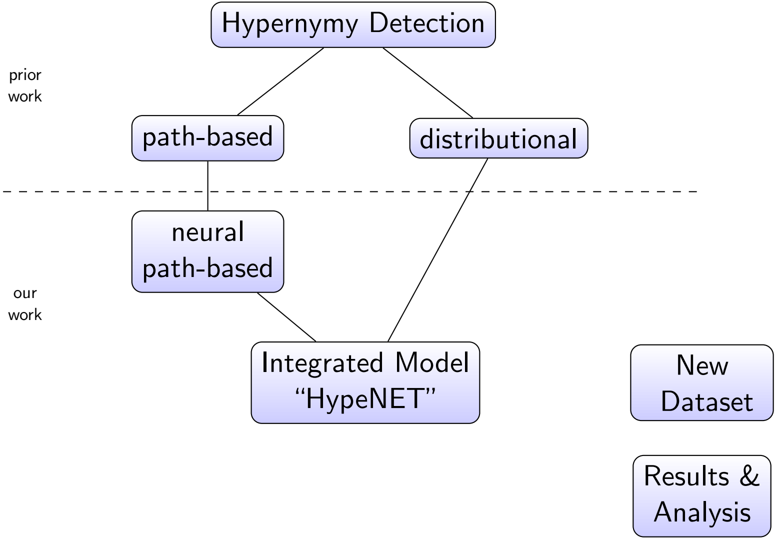
Corpus-based Methods



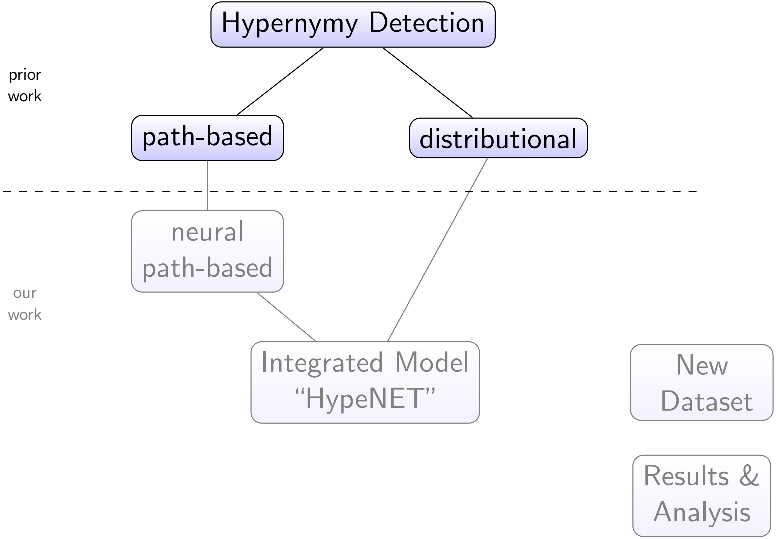
Corpus-based Methods



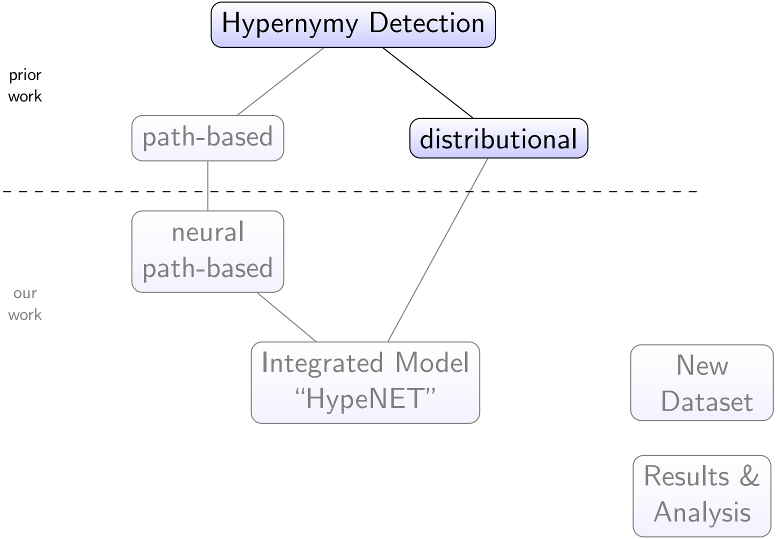
Corpus-based Methods



Prior Methods



Distributional Approach



Distributional Approach

- Recognize the relation between x and y based on their *separate* occurrences in the corpus

Distributional Approach

- Recognize the relation between x and y based on their *separate* occurrences in the corpus
- Distributional Hypothesis [Harris, 1954]:
  Words that occur in similar contexts tend to have similar meanings

Distributional Approach

- Recognize the relation between x and y based on their *separate* occurrences in the corpus

- Distributional Hypothesis [Harris, 1954]:
  Words that occur in similar contexts tend to have similar meanings
  - e.g. *elevator* and *lift* will both appear next to *up*, *floor* and *stairs*

Unsupervised Distributional Methods

- Words are represented as distributional (count-based) vectors:

0	0	...	0.56	0	...	0.89	0	...	0

$\uparrow$ up $\uparrow$ stairs

Unsupervised Distributional Methods

- Words are represented as distributional (count-based) vectors:

0	0	...	0.56	0	...	0.89	0	...	0
			↑ up			↑ stairs			

- Vector-based measures for "directional similarity":
  - Inclusion: If $x \to y$, then the prominent contexts of x are included in y [Weeds and Weir, 2003, Kotlerman et al., 2010].
  - Generality: If $x \to y$, then the most typical linguistic contexts of y are less informative than those of x [Santus et al., 2014, Rimell, 2014].

Supervised Distributional Methods

- Words are represented using low-dimensional word embeddings [Mikolov et al., 2013, Pennington et al., 2014]

Supervised Distributional Methods

- Words are represented using low-dimensional word embeddings [Mikolov et al., 2013, Pennington et al., 2014]
- (x, y) term-pairs are represented as a feature vector, based of the terms' embeddings:
  - Concatenation $\vec{x} \oplus \vec{y}$ [Baroni et al., 2012]
  - Difference $\vec{y} - \vec{x}$ [Roller et al., 2014, Weeds et al., 2014]

Supervised Distributional Methods

- Words are represented using low-dimensional word embeddings [Mikolov et al., 2013, Pennington et al., 2014]
- (x, y) term-pairs are represented as a feature vector, based of the terms' embeddings:
  - Concatenation $\vec{x} \oplus \vec{y}$ [Baroni et al., 2012]
  - Difference $\vec{y} - \vec{x}$ [Roller et al., 2014, Weeds et al., 2014]
- Train a classifier to predict whether $x \to y$ (or: y is a hypernym of x)

Supervised Distributional Methods

- Words are represented using low-dimensional word embeddings [Mikolov et al., 2013, Pennington et al., 2014]
- (x, y) term-pairs are represented as a feature vector, based of the terms' embeddings:
  - Concatenation $\vec{x} \oplus \vec{y}$ [Baroni et al., 2012]
  - Difference $\vec{y} - \vec{x}$ [Roller et al., 2014, Weeds et al., 2014]
- Train a classifier to predict whether $x \to y$ (or: y is a hypernym of x)
- Achieved very good results on common datasets

Supervised Distributional Methods

- Words are represented using low-dimensional word embeddings [Mikolov et al., 2013, Pennington et al., 2014]
- (x, y) term-pairs are represented as a feature vector, based of the terms' embeddings:
  - Concatenation $\vec{x} \oplus \vec{y}$ [Baroni et al., 2012]
  - Difference $\vec{y} - \vec{x}$ [Roller et al., 2014, Weeds et al., 2014]
- Train a classifier to predict whether $x \to y$ (or: y is a hypernym of x)

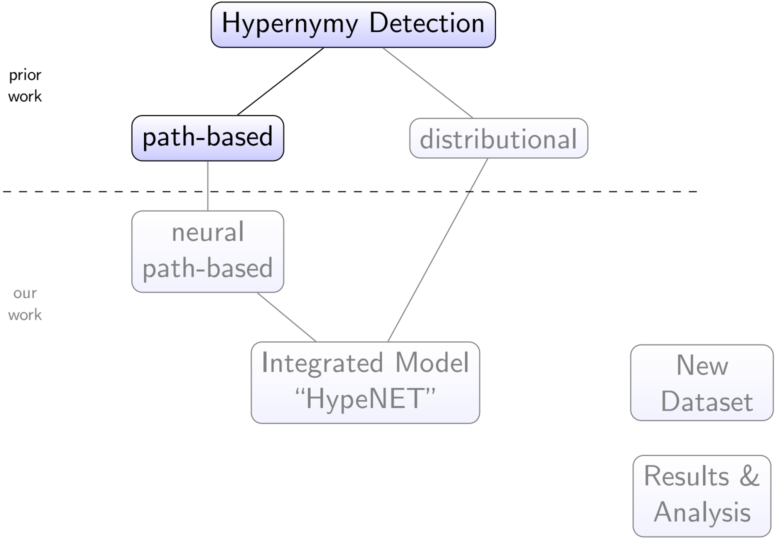
- Achieved very good results on common datasets
- Is it a solved task?

Supervised Distributional Methods

- Words are represented using low-dimensional word embeddings [Mikolov et al., 2013, Pennington et al., 2014]
- (x, y) term-pairs are represented as a feature vector, based of the terms' embeddings:
  - Concatenation $\vec{x} \oplus \vec{y}$ [Baroni et al., 2012]
  - Difference $\vec{y} - \vec{x}$ [Roller et al., 2014, Weeds et al., 2014]
- Train a classifier to predict whether $x \rightarrow y$ (or: y is a hypernym of x)

- Achieved very good results on common datasets
- Is it a solved task?
- Probably not. They don't learn the *relation* between x and y, but mostly that y is a *prototypical hypernym* [Levy et al., 2015].
  - e.g. that *(x, fruit)* or *(x, animal)* are always hypernyms

Path-based Approach



Path-based Approach

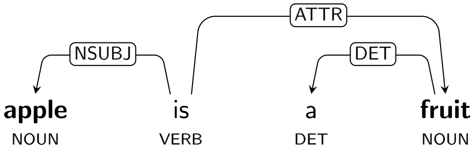
- Recognize the relation between x and y based on their *joint* occurrences in the corpus

Path-based Approach

- Recognize the relation between x and y based on their *joint* occurrences in the corpus
- Hearst Patterns [Hearst, 1992] - patterns connecting x and y may indicate that y is a hypernym of x
  - e.g. *X or other Y*, *X is a Y*, *Y, including X*

Path-based Approach

- Recognize the relation between x and y based on their *joint* occurrences in the corpus
- Hearst Patterns [Hearst, 1992] - patterns connecting x and y may indicate that y is a hypernym of x
  - e.g. *X or other Y*, *X is a Y*, *Y, including X*
- Patterns can be represented using dependency paths:



Supervised Path-based Approach

- Supervised method to recognize hypernymy [Snow et al., 2004]:

Supervised Path-based Approach

- Supervised method to recognize hypernymy [Snow et al., 2004]:
  - Features: all dependency paths that connected x and y in a corpus:

0	0	...	58	0	...	97	0	...	0

X and other Y   such Y as X

Supervised Path-based Approach

- Supervised method to recognize hypernymy [Snow et al., 2004]:
  - Features: all dependency paths that connected x and y in a corpus:

0	0	...	58	0	...	97	0	...	0
			↑ X and other Y			↑ such Y as X			

  - Supervision: set of known hyponym/hypernym pairs from WordNet

Supervised Path-based Approach

- Supervised method to recognize hypernymy [Snow et al., 2004]:
  - Features: all dependency paths that connected x and y in a corpus:

0	0	...	58	0	...	97	0	...	0
			↑ X and other Y			↑ such Y as X			

  - Supervision: set of known hyponym/hypernym pairs from WordNet
  - Trained a logistic regression classifier to predict hypernymy

Path-based Approach Issues

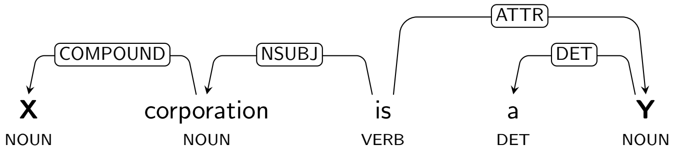
- The feature space is too sparse:

Path-based Approach Issues

- The feature space is too sparse:
    - Similar paths share no information:
      X inc. is a Y
      X group is a Y
      X organization is a Y

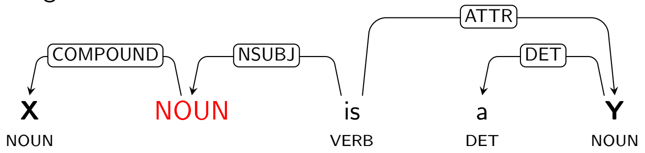
Path-based Approach Issues

- The feature space is too sparse:
  - Similar paths share no information:
    X inc. is a Y
    X group is a Y
    X organization is a Y
- PATTY [Nakashole et al., 2012] generalized paths, by replacing a word by:



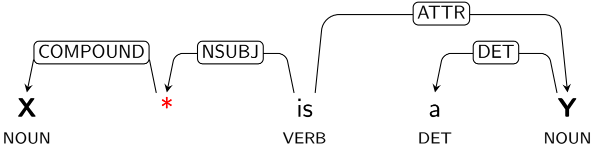
Path-based Approach Issues

- The feature space is too sparse:
  - Similar paths share no information:
    X inc. is a Y
    X group is a Y
    X organization is a Y
- PATTY [Nakashole et al., 2012] generalized paths, by replacing a word by:
  - its POS tag



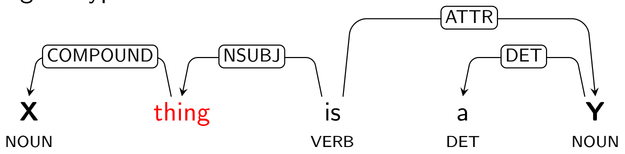
Path-based Approach Issues

- The feature space is too sparse:
  - Similar paths share no information:
    X inc. is a Y
    X group is a Y
    X organization is a Y
- PATTY [Nakashole et al., 2012] generalized paths, by replacing a word by:
  - a wild-card



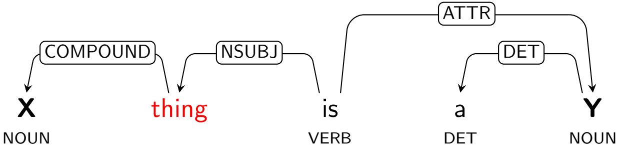
Path-based Approach Issues

- The feature space is too sparse:
  - Similar paths share no information:
    X inc. is a Y
    X group is a Y
    X organization is a Y
- PATTY [Nakashole et al., 2012] generalized paths, by replacing a word by:
  - its ontological type



Path-based Approach Issues

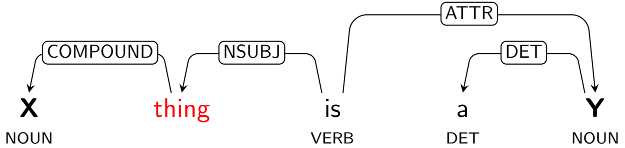
- The feature space is too sparse:
  - Similar paths share no information:
    X inc. is a Y
    X group is a Y
    X organization is a Y
- PATTY [Nakashole et al., 2012] generalized paths, by replacing a word by:
  - its ontological type



- Some of these generalizations are too general:
  - X is defined as Y ≈ X is described as Y via X is VERB as Y

Path-based Approach Issues

- The feature space is too sparse:
  - Similar paths share no information:
    X inc. is a Y
    X group is a Y
    X organization is a Y
- PATTY [Nakashole et al., 2012] generalized paths, by replacing a word by:
  - its ontological type



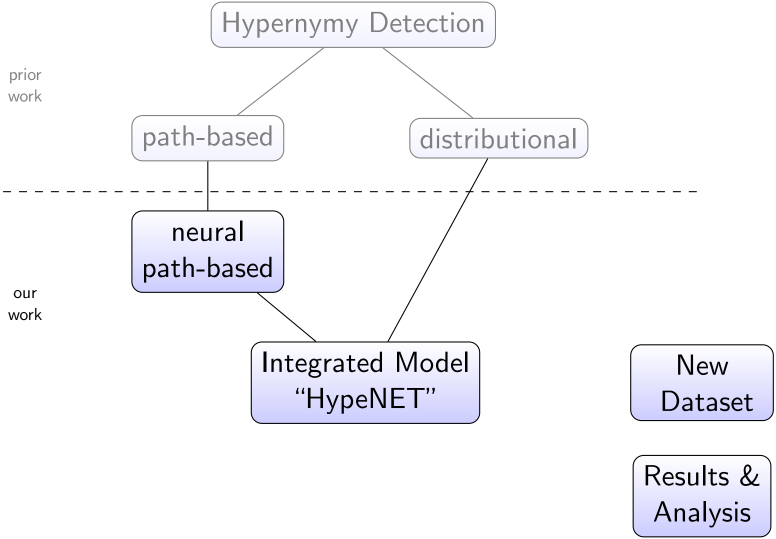
- Some of these generalizations are too general:
  - X is defined as Y $\approx$ X is described as Y via X is VERB as Y
  - X is defined as Y $\neq$ X is rejected as Y

Improving Hypernymy Detection with an Integrated Path-based and Distributional Method

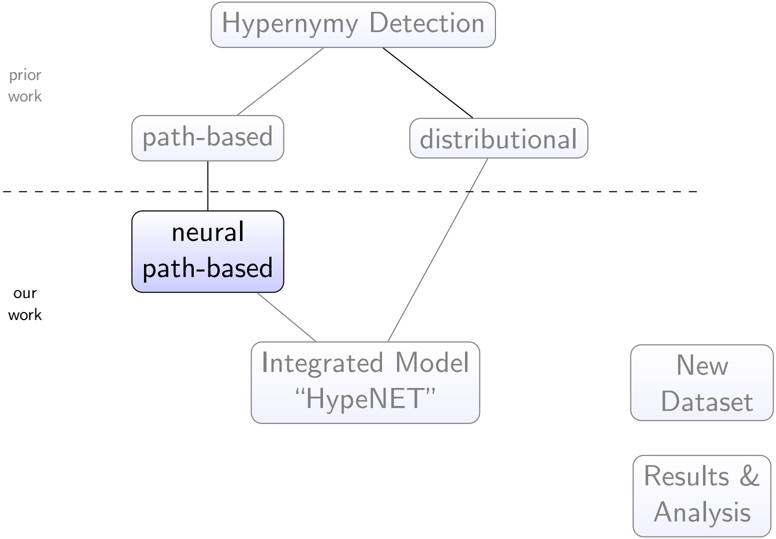
Vered Shwartz, Yoav Goldberg and Ido Dagan

ACL 2016

HypeNET: Integrated Path-based and Distributional Method



First Step: Improving Path Representation



Path Representation (1/2)

1. Split each path to edges

X	is	a	Y	⇒
`‘X/NOUN/nsubj/>`	`be/VERB/ROOT/-`	`Y/NOUN/attr/<’`		⇒
`‘X/NOUN/nsubj/>’`	`‘be/VERB/ROOT/-’`	`‘Y/NOUN/attr/<’`		

- Each edge consists of 4 components:

  `dependent lemma` / `dependent POS` / `dependency label` / `direction`

Path Representation (1/2)

1. Split each path to edges

X is a Y ⇒
'X/NOUN/nsubj/> be/VERB/ROOT/- Y/NOUN/attr/<' ⇒
'X/NOUN/nsubj/>' 'be/VERB/ROOT/-' 'Y/NOUN/attr/<'

- Each edge consists of 4 components:

  `dependent lemma` / `dependent POS` / `dependency label` / `direction`

- We learn embedding vectors for each component
  - Lemma embeddings are initialized with pre-trained word embeddings

Path Representation (1/2)

1. Split each path to edges

   X is a Y ⇒

   'X/NOUN/nsubj/> be/VERB/ROOT/- Y/NOUN/attr/<' ⇒

   'X/NOUN/nsubj/>' 'be/VERB/ROOT/-' 'Y/NOUN/attr/<'

   - Each edge consists of 4 components:

     dependent lemma / dependent POS / dependency label / direction

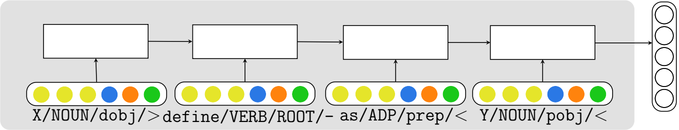
   - We learn embedding vectors for each component
     - Lemma embeddings are initialized with pre-trained word embeddings
   - The edge's vector is the concatenation of its components' vectors:

be/VERB/ROOT/-

   - Generalization: similar edges should have similar vectors!

Path Representation (2/2)

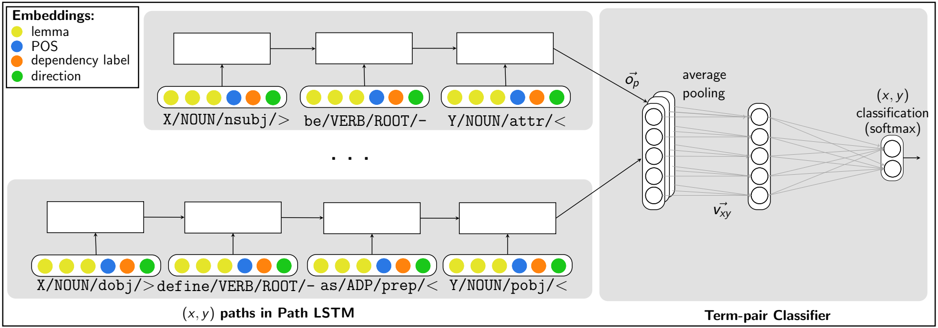
2. Feed the edges sequentially to an LSTM



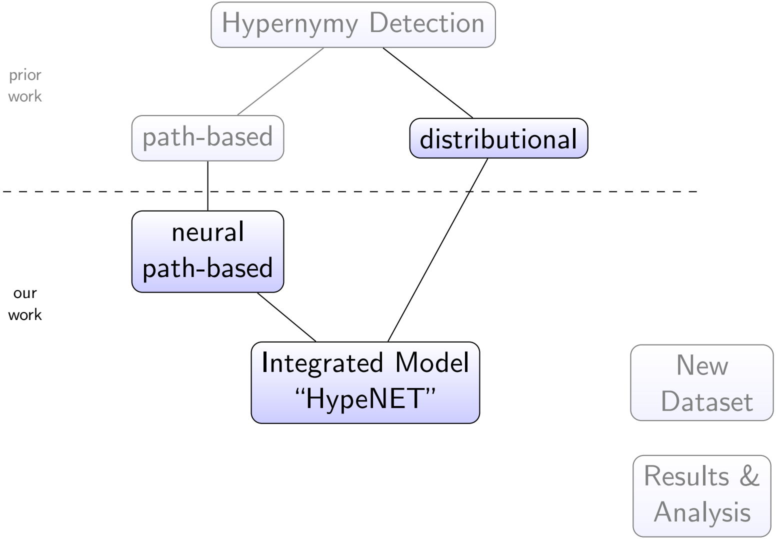
- Use the last output vector as the path embedding
- The LSTM may focus on edges that are more informative for the classification task, while ignoring others

Term-pair Classification

- The LSTM encodes a single path
- Each term-pair has multiple paths
  - Represent a term-pair as its averaged path embedding
- Classify for hypernymy (path-based network):

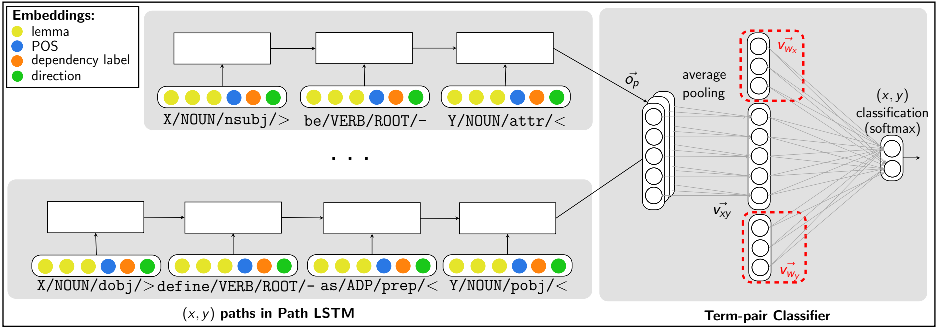


Second Step: Integrating Distributional Information

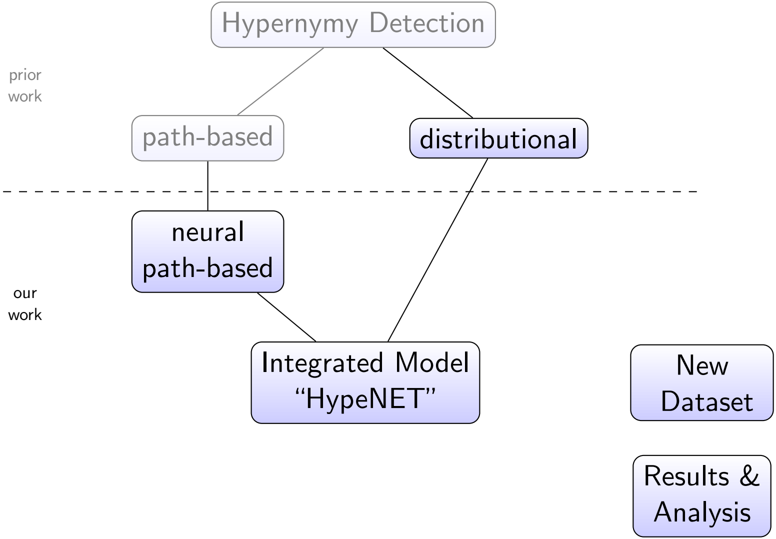


Second Step: Integrating Distributional Information

- Integrated network: add distributional information
  - Simply concatenate x and y's word embeddings to the averaged path
- Classify for hypernymy (integrated network):



Evaluation



Hypernymy Detection
prior
work
path-based
distributional
neural
path-based
our
work
Integrated Model
"HypeNET"
New
Dataset
Results &
Analysis

New Dataset

- Distant supervision from knowledge resources
    - Size: 70,679 entries
    - Positive instances: term-pairs related via hypernymy relations
        - e.g. *instance_of*
    - Negative instances: term-pairs related via other relations
    - Filtering: pairs must co-occur at least twice (like [Snow et al., 2004])

New Dataset

- Distant supervision from knowledge resources
  - Size: 70,679 entries
  - Positive instances: term-pairs related via hypernymy relations
    - e.g. *instance_of*
  - Negative instances: term-pairs related via other relations
  - Filtering: pairs must co-occur at least twice (like [Snow et al., 2004])
- Train / test / validation split: random (70% - 25% - 5%)

Results

method		precision	recall	F_1
Path-based	Snow	0.843	0.452	0.589
	Snow + GEN	0.852	0.561	0.676
	HypeNET Path-based	0.811	0.716	0.761
Distributional	Best Supervised	0.901	0.637	0.746
Combined	HypeNET Integrated	**0.913**	**0.890**	**0.901**

- Path-based:
  - Compared to Snow + Snow with PATTY style generalizations
  - Our method outperforms path-based baselines with improved recall

Results

method		precision	recall	F_1
Path-based	Snow	0.843	0.452	0.589
	Snow + GEN	0.852	0.561	0.676
	HypeNET Path-based	0.811	0.716	0.761
Distributional	Best Supervised	0.901	0.637	0.746
Combined	HypeNET Integrated	**0.913**	**0.890**	**0.901**

- Distributional:
  - Compared to several supervised/unsupervised methods
  - HypeNET Path-based performs similarly to best distributional method

Results

method		precision	recall	F_1
Path-based	Snow	0.843	0.452	0.589
	Snow + GEN	0.852	0.561	0.676
	HypeNET Path-based	0.811	0.716	0.761
Distributional	Best Supervised	0.901	0.637	0.746
Combined	HypeNET Integrated	**0.913**	**0.890**	**0.901**

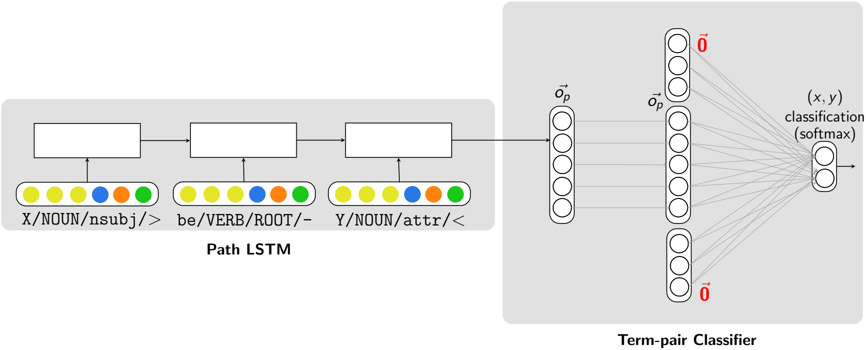
- The integrated method substantially outperforms both path-based and distributional methods

Analysis

Analysis - Path Representation (1/2)

- Identify hypernymy-indicating paths:
  - Baselines: according to logistic regression feature weights

- Identify hypernymy-indicating paths:
  - Baselines: according to logistic regression feature weights
  - HypeNET: measure path contribution to positive classification:



- Take the top scoring paths according to $softmax(W \cdot [\vec{0}, \vec{o_p}, \vec{0}])[1]$

Analysis - Path Representation (2/2)

- Snow's method finds certain common paths:

  X company is a Y

  X ltd is a Y

Analysis - Path Representation (2/2)

- Snow's method finds certain common paths:

  X company is a Y

  X ltd is a Y

- PATTY-style generalizations find very general, possibly noisy paths:

  X NOUN is a Y

Analysis - Path Representation (2/2)

- Snow's method finds certain common paths:

  X company is a Y
  X ltd is a Y

- PATTY-style generalizations find very general, possibly noisy paths:

  X NOUN is a Y

- HypeNET makes fine-grained generalizations:

  X association is a Y
  X co. is a Y
  X company is a Y
  X corporation is a Y
  X foundation is a Y
  X group is a Y
  ...

What's Next?

What's Next?

- Recognizing lexical inferences within context:

What's Next?

- Recognizing lexical inferences within context:

  - Detect the correct sense of a term (*apple*) within the given context

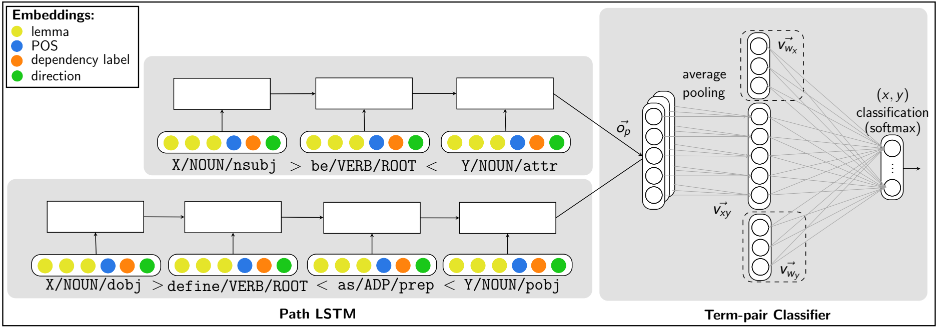
What's Next?

- Recognizing lexical inferences within context:

  - Detect the correct sense of a term (*apple*) within the given context
  - Base the entailment decision on the sentence and the semantic relation, e.g. hypernymy in upward/downward monotone sentences:
    - I ate an *apple* $\rightarrow$ I ate a *fruit*
    - I hate *fruit* $\rightarrow$ I hate *apples*

First Step: Recognizing Multiple Semantic Relations I

1. Extend HypeNET to support multiple semantic relations (LexNET):



First Step: Recognizing Multiple Semantic Relations II

2. Test on common semantic relations datasets:

	K&H+N			BLESS			ROOT09			EVALution		
method	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
Path-based	0.713	0.604	0.55	0.759	0.756	0.755	0.788	0.789	0.788	0.53	0.537	0.503
Distributional	0.909	0.906	0.904	0.811	0.812	0.811	0.636	0.675	0.646	0.531	0.544	0.525
Distributional NN	0.983	0.984	0.983	0.891	0.889	0.889	0.712	0.721	0.716	0.57	0.573	0.571
LexNET	**0.985**	**0.986**	**0.985**	**0.894**	**0.893**	**0.893**	**0.813**	**0.814**	**0.813**	**0.601**	**0.607**	**0.6**

- LexNET outperforms individual path-based and distributional methods
- Path-based contribution over distributional info small but consistent, especially when:
  - The dataset is not biased (i.e. when lexical memorization is disabled), e.g. *random:(toaster, vehicle)*.
  - x or y are polysemous, e.g. *mero:(piano, key)*.
  - x or y are rare, e.g. *hyper:(mastodon, proboscidean)*.

First Step: Recognizing Multiple Semantic Relations III

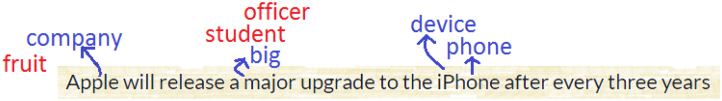
3. Which relations can be learned? ($F_1 \times 100$ score):

relation	Path-based	Distributional	LexNET	Δ
antonym	23.6	50.7	52.9	+2.2
attribute	78.0	85.9	88.4	+2.5
co-hyponym	37.9	94.4	96.2	+1.8
event	73.9	88.3	89.3	+1.0
hypernym	57.8	80.4	81.3	+0.9
meronym	60.4	80.1	82.3	+2.2
synonym	10.6	33.3	36.4	+3.1

- Both methods are not good in recognizing synonyms and antonyms.
- Path-based information adds over the distributional one mostly for relations that it learned well (attribute) or for those that the distributional method did not (synonym).

Future Work: Lexical Inference in Context

- Developing a method to recognize lexical inferences within context:



Questions?

References

Baroni, M., Bernardi, R., Do, N.-Q., and Shan, C.-c. (2012).
Entailment above the word level in distributional semantics.
In *EACL*, pages 23–32.

Harris, Z. S. (1954).
Distributional structure.
Word, 10(2-3):146–162.

Hearst, M. A. (1992).
Automatic acquisition of hyponyms from large text corpora.
In *ACL*, pages 539–545.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. (2010).
Directional distributional similarity for lexical inference.
NLE, 16(04):359–389.

Levy, O., Remus, S., Biemann, C., and Dagan, I. (2015).
Do supervised distributional methods really learn lexical inference relations.
NAACL.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In *NIPS*, pages 3111–3119.

Nakashole, N., Weikum, G., and Suchanek, F. (2012).
Patty: a taxonomy of relational patterns with semantic types.
In *EMNLP and CoNLL*, pages 1135–1145.

Pennington, J., Socher, R., and Manning, C. D. (2014).
Glove: Global vectors for word representation.
In *EMNLP*, pages 1532–1543.

Rimell, L. (2014).
Distributional lexical entailment by topic coherence.
In *EACL*, pages 511–519.

Roller, S., Erk, K., and Boleda, G. (2014).
Inclusive yet selective: Supervised distributional hypernymy detection.
In *COLING*, pages 1025–1036.

Santus, E., Lenci, A., Lu, Q., and Im Walde, S. S. (2014).
Chasing hypernyms in vector spaces with entropy.
In *EACL*, pages 38–42.

Snow, R., Jurafsky, D., and Ng, A. Y. (2004).
Learning syntactic patterns for automatic hypernym discovery.
In *NIPS*.

Weeds, J., Clarke, D., Reffin, J., Weir, D., and Keller, B. (2014).
Learning to distinguish hypernyms and co-hyponyms.
In *COLING*, pages 2249–2259.

Weeds, J. and Weir, D. (2003).
A general framework for distributional similarity.
In *EMLP*, pages 81–88.