Hypernyms under Siege:
Linguistically-motivated Artillery for Hypernymy Detection

Vered Shwartz\textsuperscript{1}, Enrico Santus\textsuperscript{2} and Dominik Schlechtweg\textsuperscript{3}

\textsuperscript{1}Bar-Ilan University, \textsuperscript{2}Singapore University of Technology and Design, \textsuperscript{3}University of Stuttgart

EACL 2017
Hypernymy Detection - Task Definition

- Given two terms, \((x, y)\), decide whether \(y\) is a hypernym of \(x\)

  e.g. Python's hypernyms are animal and programming language

Motivation: taxonomy creation, RTE, QA, ...
Hypernymy Detection - Task Definition

- Given two terms, \((x, y)\), decide whether \(y\) is a hypernym of \(x\)
  - in some senses of \(x\) and \(y\)
Hypernymy Detection - Task Definition

- Given two terms, \((x, y)\), decide whether \(y\) is a hypernym of \(x\)
  - in some senses of \(x\) and \(y\)
  - e.g. `python`'s hypernyms are `animal` and `programming language`
Hypernymy Detection - Task Definition

Given two terms, \((x, y)\), decide whether \(y\) is a hypernym of \(x\)

- in some senses of \(x\) and \(y\)
- e.g. *python*'s hypernyms are *animal* and *programming language*

Motivation: taxonomy creation, RTE, QA, ...
Hypernymy Detection - Task Definition

Given two terms, \((x, y)\), decide whether \(y\) is a hypernym of \(x\)
- in some senses of \(x\) and \(y\)
- e.g. python’s hypernyms are animal and programming language

Motivation: taxonomy creation, RTE, QA, ...
- e.g.: A boy is hitting a baseball \(\Rightarrow\) A child is hitting a baseball
Hypernymy Detection

**corpus-based**

- **path-based**
  - Hearst Patterns [Hearst, 1992]
  - Snow [Snow et al., 2005]
  - Neculescu et al., 2015
  - ...

- **distributional**
  - HypeNET [Shwartz et al., 2016]

**unsupervised**

- **similarity**
  - Cosine [Salton and McGill, 1986]
  - Lin [Lin, 1998]
  - APSyn [Santus et al., 2016b]

- **informativeness**
  - SLQS [Santus et al., 2014]
  - RCTC [Rimell, 2014]

- **inclusion**
  - WeedsPrec [Weeds and Weir, 2003]
  - ClarkeDE [Clarke, 2009]
  - balAPinc [Kotlerman et al., 2010]
  - ...

**supervised**

- concat [Baroni et al., 2012]
- difference [Weeds et al., 2014]
- ASYM [Roller et al., 2014]
- ...
Unsupervised Hypernymy Detection

**corpus-based**
- path-based
  - Hearst Patterns [Hearst, 1992]
    - Snow [Snow et al., 2005]
    - [Necsulescu et al., 2015]
  - HypeNET [Shwartz et al., 2016]

**distributional**
- unsupervised
  - similarity
    - Cosine [Salton and McGill, 1986]
    - Lin [Lin, 1998]
    - APSyn [Santus et al., 2016b]
  - informativeness
    - SLQS [Santus et al., 2014]
    - RCTC [Rimell, 2014]
  - inclusion
    - WeedsPrec [Weeds and Weir, 2003]
    - ClarkeDE [Clarke, 2009]
    - balAPinc [Kotlerman et al., 2010]

**supervised**
- concat [Baroni et al., 2012]
- difference [Weeds et al., 2014]
- ASYM [Roller et al., 2014]
Key Takeouts

- We performed an extensive number of experiments, with:
  - 14 unsupervised measures
  - 12 DSMs: 6 context types $\times$ 2 feature weightings
  - 4 datasets
Key Takeouts

- We performed an extensive number of experiments, with:
  - 14 unsupervised measures
  - 12 DSMs: 6 context types $\times$ 2 feature weightings
  - 4 datasets

- Our findings:
  1. There is no specific measure which is always preferred

Different measures capture different aspects of hypernymy.
Specific measures are preferred in distinguishing hypernymy from each other relation.

Unsupervised measures are not deprecated given supervised methods.
In real-application scenario, unsupervised may be preferred.
Key Takeouts

- We performed an extensive number of experiments, with:
  - 14 unsupervised measures
  - 12 DSMs: 6 context types × 2 feature weightings
  - 4 datasets

- Our findings:
  1. There is no specific measure which is always preferred
     - Different measures capture different aspects of hypernymy
Key Takeouts

- We performed an extensive number of experiments, with:
  - 14 unsupervised measures
  - 12 DSMs: 6 context types × 2 feature weightings
  - 4 datasets

- Our findings:
  1. There is no specific measure which is always preferred
     - Different measures capture different aspects of hypernymy
     - Specific measures are preferred in distinguishing hypernymy from each other relation
Key Takeouts

- We performed an extensive number of experiments, with:
  - 14 unsupervised measures
  - 12 DSMs: 6 context types × 2 feature weightings
  - 4 datasets

- Our findings:
  1. There is no specific measure which is always preferred
     - Different measures capture different aspects of hypernymy
     - Specific measures are preferred in distinguishing hypernymy from each other relation
  2. Unsupervised measures are not deprecated given supervised methods
Key Takeouts

- We performed an extensive number of experiments, with:
  - 14 unsupervised measures
  - 12 DSMs: 6 context types $\times$ 2 feature weightings
  - 4 datasets

- Our findings:
  1. There is no specific measure which is always preferred
     - Different measures capture different aspects of hypernymy
     - Specific measures are preferred in distinguishing hypernymy from each other relation

  2. Unsupervised measures are not deprecated given supervised methods
     - In real-application scenario, unsupervised may be preferred
Outline

Comparing Between Unsupervised Measures

Hypernym vs. All Other Relations
Hypernym vs. One Other Relation

Supervised vs. Unsupervised

Performance
Robustness
Outline 1

Comparing Between Unsupervised Measures

Hypernym vs. All Other Relations
Hypernym vs. One Other Relation

Supervised vs. Unsupervised

Performance
Robustness
Hyponym vs. All

Q: Is there a “best” unsupervised measure?
Q: Is there a “best” unsupervised measure?
Evaluation setting: hypernymy vs. all other relations
Q: Is there a “best” unsupervised measure?
Evaluation setting: hypernymy vs. all other relations
Compute measure score for each \((x, y)\) pair

<table>
<thead>
<tr>
<th>Measure</th>
<th>Context Type</th>
<th>AP@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVALution</td>
<td>invCL</td>
<td>0.661</td>
</tr>
<tr>
<td>BLESS</td>
<td>invCL</td>
<td>0.54</td>
</tr>
<tr>
<td>Lenci/Benotto</td>
<td>APSyn</td>
<td>0.617</td>
</tr>
<tr>
<td>Weeds</td>
<td>clarkeDE</td>
<td>0.911</td>
</tr>
</tbody>
</table>
Hypernym vs. All

- Q: Is there a “best” unsupervised measure?
- Evaluation setting: hypernymy vs. all other relations
- Compute measure score for each \((x, y)\) pair
- Evaluation metric: average precision
Q: Is there a “best” unsupervised measure?
Evaluation setting: hypernymy vs. all other relations
Compute measure score for each \((x, y)\) pair
Evaluation metric: average precision
A: There is no single combination of measure, context type and feature weighting that consistently performs best

<table>
<thead>
<tr>
<th>dataset</th>
<th>measure</th>
<th>context type</th>
<th>AP@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVALution</td>
<td>invCL</td>
<td>joint</td>
<td>0.661</td>
</tr>
<tr>
<td>BLESS</td>
<td>invCL</td>
<td>win5</td>
<td>0.54</td>
</tr>
<tr>
<td>Lenci/Benotto</td>
<td>APSyn</td>
<td>joint</td>
<td>0.617</td>
</tr>
<tr>
<td>Weeds</td>
<td>clarkeDE</td>
<td>win5d</td>
<td>0.911</td>
</tr>
</tbody>
</table>
Hypernym vs. One Other Relation

- Different measures capture different aspects of hypernymy
  - e.g., that $x$’s contexts are included in $y$’s
Different measures capture different aspects of hypernymy
- e.g., that $x$'s contexts are included in $y$'s

Q: Is there a most suitable measure to distinguish hypernymy from each other relation?
Different measures capture different aspects of hypernymy
- e.g., that x’s contexts are included in y’s

Q: Is there a most suitable measure to distinguish hypernymy from each other relation?

Consider all relations that occurred in two datasets
Different measures capture different aspects of hypernymy
  e.g., that $x$'s contexts are included in $y$'s

Q: Is there a most suitable measure to distinguish hypernymy from each other relation?

Consider all relations that occurred in two datasets

For such relation, for each dataset, rank measures by AP and select those with AP $\geq 0.8$
Hypernym vs. Attribute

- Best measures: similarity measures + syntactic contexts
Hypernym vs. Attribute

- Best measures: similarity measures + syntactic contexts

Possible explanation:
- Similarity measures quantify the ratio of shared contexts between $x$ and $y$
Hypernym vs. Attribute

- Best measures: similarity measures + syntactic contexts

Possible explanation:

- Similarity measures quantify the ratio of shared contexts between \( x \) and \( y \)
- Attributes contexts’ are easy to distinguish syntactically
Hypernym vs. Attribute

- Best measures: similarity measures + syntactic contexts

Possible explanation:
- Similarity measures quantify the ratio of shared contexts between $x$ and $y$
- Attributes contexts’ are easy to distinguish syntactically
- $y$ is almost always an adjective, $x$ is a noun/verb
Hypernym vs. Attribute

- Best measures: similarity measures + syntactic contexts

Possible explanation:
- Similarity measures quantify the ratio of shared contexts between $x$ and $y$
- Attributes contexts’ are easy to distinguish syntactically
- $y$ is almost always an adjective, $x$ is a noun/verb
- Less shared contexts than in hyponym-hypernym pairs
Hypernym vs. Meronym

- Best measures: inclusion measures + syntactic contexts

- Possible explanation:
  - Inclusion measures quantify the ratio of $x$'s contexts included in $y$'s [Weeds and Weir, 2003, Geffet and Dagan, 2005]
  - Window-based contexts capture topical context
  - Topics are shared between holonym-meronym and hypernym-hyponym pairs
  - Syntactic contexts capture functional context
  - Meronyms and holonyms are more often of different functions than hyponym and hypernyms
  - *cat* and *animal* both eat and are alive
  - *you* wag a *tail* but not a *cat*
Hypernym vs. Meronym

- Best measures: inclusion measures + syntactic contexts

Possible explanation:
- Inclusion measures quantify the ratio of $x$'s contexts included in $y$'s [Weeds and Weir, 2003, Geffet and Dagan, 2005]
Hypernym vs. Meronym

- Best measures: inclusion measures + syntactic contexts

Possible explanation:

- Inclusion measures quantify the ratio of $x$’s contexts included in $y$’s [Weeds and Weir, 2003, Geffet and Dagan, 2005]
- Window-based contexts capture *topical* context
Hyponym vs. Meronym

- Best measures: inclusion measures + syntactic contexts

Possible explanation:
- Inclusion measures quantify the ratio of $x$'s contexts included in $y$'s [Weeds and Weir, 2003, Geffet and Dagan, 2005]
- Window-based contexts capture *topical* context
- Topics are shared between holonym-meronym and hyponym-hyponym pairs
Hypernym vs. Meronym

- Best measures: inclusion measures + syntactic contexts

Possible explanation:

- Inclusion measures quantify the ratio of $x$’s contexts included in $y$’s [Weeds and Weir, 2003, Geffet and Dagan, 2005]
- Window-based contexts capture *topical* context
- Topics are shared between holonym-meronym and hypernym-hyponym pairs
- Syntactic contexts capture *functional* context
Hypernym vs. Meronym

Best measures: inclusion measures + syntactic contexts

Possible explanation:

- Inclusion measures quantify the ratio of $x$'s contexts included in $y$’s [Weeds and Weir, 2003, Geffet and Dagan, 2005]
- Window-based contexts capture *topical* context
- Topics are shared between holonym-meronym and hypernym-hyponym pairs
- Syntactic contexts capture *functional* context
- Meronyms and holonyms are more often of different functions than hyponym and hypernyms

*cat* and *animal* both eat and are alive

*you* wag a *tail* but not a *cat*
Hypernym vs. Meronym

- Best measures: inclusion measures + syntactic contexts

Possible explanation:

- Inclusion measures quantify the ratio of $x$’s contexts included in $y$’s [Weeds and Weir, 2003, Geffet and Dagan, 2005]
- Window-based contexts capture *topical* context
- Topics are shared between holonym-meronym and hypernym-hyponym pairs
- Syntactic contexts capture *functional* context
- Meronyms and holonyms are more often of different functions than hyponym and hypernyms
  - *cat* and *animal* both eat and are alive
  - you wag a *tail* but not a *cat*
Hypernym vs. Synonym (1/2)

- Best measures: SLQS / invCL

Possible explanation:

- SLQS [Santus et al., 2014]: $x$ is more informative than $y$.
- $y$ is less informative than $x$ in hypernymy, not in synonymy.
- Budgie is more informative than bird.
- Budgie and parakeet are of the same informativeness level.
Hypernym vs. Synonym (1/2)

- Best measures: SLQS / invCL

Possible explanation:
- **SLQS** [Santus et al., 2014]: is $x$ more informative than $y$?

**budgie** is more informative than **bird**

**budgie** and **parakeet** are of the same informativeness level
Hypernym vs. Synonym (1/2)

- Best measures: SLQS / invCL

Possible explanation:
- **SLQS** [Santus et al., 2014]: is $x$ more informative than $y$?
- $y$ is less informative than $x$ in hypernymy, not in synonymy
Hypernym vs. Synonym (1/2)

- Best measures: SLQS / invCL

Possible explanation:

- **SLQS** [Santus et al., 2014]: is $x$ more informative than $y$?
- $y$ is less informative than $x$ in hypernymy, not in synonymy
  - *budgie* is more informative than *bird*
  - *budgie* and *parakeet* are of the same informativeness level
Hypernym vs. Synonym (2/2)

- **invCL** [Lenci and Benotto, 2012] is a *strict* inclusion measure, quantifying also the non-inclusion of $y$ in $x$. Inclusion measures may assign high scores to synonyms, since they share many contexts. **invCL** penalizes $(x, y)$ pairs in which $x$ also includes many of $y$'s contexts.
Hypernym vs. Synonym (2/2)

- **invCL** [Lenci and Benotto, 2012] is a *strict* inclusion measure, quantifying also the non-inclusion of \( y \) in \( x \).
- Inclusion measures may assign high scores to synonyms, since they share many contexts.
invCL [Lenci and Benotto, 2012] is a strict inclusion measure, quantifying also the non-inclusion of \( y \) in \( x \)

Inclusion measures may assign high scores to synonyms, since they share many contexts

invCL penalizes \((x, y)\) pairs in which \( x \) also includes many of \( y \)'s contexts
Comparing Between Unsupervised Measures

- Hypernym vs. All Other Relations
- Hypernym vs. One Other Relation

Supervised vs. Unsupervised

- Performance
- Robustness
Supervised Methods

- (Distributional) state-of-the-art: supervised embedding-based methods

- Glide
  - Concatenation: $\vec{x} \oplus \vec{y}$ [Baroni et al., 2012]
  - Difference: $\vec{y} - \vec{x}$ [Weeds et al., 2014]
  - ASYM [Roller et al., 2014]

Evaluation setting:
- Tune hyper-parameters (methods, pre-trained embeddings)
- Train a logistic regression classifier
- Use logistic score for ranking, report AP
Supervised Methods

- (Distributional) state-of-the-art: supervised embedding-based methods

- \((x, y)\) term-pairs are represented as a feature vector, based of the terms’ embeddings:
  - Concatenation \(\vec{x} \oplus \vec{y}\) [Baroni et al., 2012]
  - Difference \(\vec{y} - \vec{x}\) [Weeds et al., 2014]
  - ASYM [Roller et al., 2014]
Supervised Methods

- (Distributional) state-of-the-art: supervised embedding-based methods

- \((x, y)\) term-pairs are represented as a feature vector, based of the terms’ embeddings:
  - Concatenation \(\vec{x} \oplus \vec{y}\) [Baroni et al., 2012]
  - Difference \(\vec{y} - \vec{x}\) [Weeds et al., 2014]
  - ASYM [Roller et al., 2014]

- Evaluation setting:
  - Tune hyper-parameters (methods, pre-trained embeddings)
  - Train a logistic regression classifier
  - Use logistic score for ranking, report AP
Supervised vs. Unsupervised Experiment

- Concat achieves almost perfect AP scores, outperforms unsupervised measures:

<table>
<thead>
<tr>
<th>dataset</th>
<th>relation</th>
<th>best supervised</th>
<th>best unsupervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVALution</td>
<td>meronym</td>
<td>0.998</td>
<td>0.886</td>
</tr>
<tr>
<td></td>
<td>attribute</td>
<td>1.000</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>antonym</td>
<td>1.000</td>
<td>0.773</td>
</tr>
<tr>
<td></td>
<td>synonym</td>
<td>0.996</td>
<td>0.813</td>
</tr>
<tr>
<td>BLESS</td>
<td>meronym</td>
<td>1.000</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>coord</td>
<td>1.000</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>attribute</td>
<td>1.000</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>event</td>
<td>1.000</td>
<td>0.847</td>
</tr>
<tr>
<td></td>
<td>random-n</td>
<td>0.995</td>
<td>0.818</td>
</tr>
<tr>
<td></td>
<td>random-j</td>
<td>1.000</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td>random-v</td>
<td>1.000</td>
<td>0.895</td>
</tr>
<tr>
<td>Lenci/Benotto</td>
<td>antonym</td>
<td>0.917</td>
<td>0.807</td>
</tr>
<tr>
<td></td>
<td>synonym</td>
<td>0.946</td>
<td>0.914</td>
</tr>
<tr>
<td>Weeds</td>
<td>coord</td>
<td>0.873</td>
<td>0.824</td>
</tr>
</tbody>
</table>
Measuring Robustness (1/3)

▶ Is there any advantage to unsupervised measures?

- Apart from not requiring training data...
- Supervised methods do not learn the relation between $x$ and $y$.
  - [Levy et al., 2015]: They memorize that $y$ is a prototypical hypernym.
  - [Roller and Erk, 2016]: They trace $y$'s occurrences in Hearst patterns.
  - [Shwartz et al., 2016]: They provide the "prior" of $y$ to fit the relation.
- AP scores may suggest overfitting by the supervised methods...
Measuring Robustness (1/3)

- Is there any advantage to unsupervised measures?
  - apart from not requiring training data...
Measuring Robustness (1/3)

- Is there any advantage to unsupervised measures?
  - apart from not requiring training data...

- Supervised methods do not learn the relation between $x$ and $y$:
Measuring Robustness (1/3)

- Is there any advantage to unsupervised measures?
  - apart from not requiring training data...

- Supervised methods do not learn the relation between $x$ and $y$:
  - [Levy et al., 2015]: they memorize that $y$ is a prototypical hypernym
Measuring Robustness (1/3)

- Is there any advantage to unsupervised measures?
  - apart from not requiring training data...

- Supervised methods do not learn the relation between \( x \) and \( y \):
  - [Levy et al., 2015]: they memorize that \( y \) is a *prototypical hypernym*
  - [Roller and Erk, 2016]: they trace \( y \)'s occurrences in Hearst patterns
Measuring Robustness (1/3)

▶ Is there any advantage to unsupervised measures?
  ▶ apart from not requiring training data...

▶ Supervised methods do not learn the relation between \( x \) and \( y \):
  ▶ [Levy et al., 2015]: they memorize that \( y \) is a prototypical hypernym
  ▶ [Roller and Erk, 2016]: they trace \( y \)'s occurrences in Hearst patterns
  ▶ [Shwartz et al., 2016]: they provide the “prior” of \( y \) to fit the relation...
Is there any advantage to unsupervised measures? 
  - apart from not requiring training data...

Supervised methods do not learn the relation between $x$ and $y$:
  - [Levy et al., 2015]: they memorize that $y$ is a *prototypical hypernym*
  - [Roller and Erk, 2016]: they trace $y$'s occurrences in Hearst patterns
  - [Shwartz et al., 2016]: they provide the “prior” of $y$ to fit the relation

AP scores may suggest overfitting by the supervised methods...
Measuring Robustness (2/3)

- We repeated the experiment of [Santus et al., 2016a]
  - Adding switched hypernym pairs as random
  - e.g. (apple, animal), (cow, fruit)

- We added 139 such new random pairs
- We used the best supervised and unsupervised methods to re-classify the revised dataset
Measuring Robustness (2/3)

- We repeated the experiment of [Santus et al., 2016a]
  - Adding switched hypernym pairs as random
  - e.g. *apple, animal*, *cow, fruit*

- Using BLESS (the only dataset with random pairs)
Measuring Robustness (2/3)

- We repeated the experiment of [Santus et al., 2016a]
  - Adding switched hypernym pairs as random
  - e.g. \((apple, animal), (cow, fruit)\)

- Using BLESS (the only dataset with random pairs)
- For each hypernym pair \((x_1, y_1)\):
  - Sampled \(y_2\) such that \((x_2, y_2) \in\) dataset is a hypernym-pair

We added 139 such new random pairs

We used the best supervised and unsupervised methods to re-classify the revised dataset
Measuring Robustness (2/3)

- We repeated the experiment of [Santus et al., 2016a]
  - Adding switched hypernym pairs as random
  - e.g. \((apple, animal), (cow, fruit)\)

- Using BLESS (the only dataset with random pairs)
- For each hypernym pair \((x_1, y_1)\):
  - Sampled \(y_2\) such that \((x_2, y_2)\) \(\in\) dataset is a hypernym-pair
  - such that \((x_1, y_2)\) \(\notin\) dataset

- We added 139 such new random pairs
- We used the best supervised and unsupervised methods to re-classify the revised dataset
Measuring Robustness (2/3)

- We repeated the experiment of [Santus et al., 2016a]
  - Adding switched hypernym pairs as random
  - e.g. (apple, animal), (cow, fruit)

- Using BLESS (the only dataset with random pairs)
- For each hypernym pair \((x_1, y_1)\):
  - Sampled \(y_2\) such that \((x_2, y_2)\) \(\in\) dataset is a hypernym-pair
  - such that \((x_1, y_2)\) \(\notin\) dataset
  - Add \((x_1, y_2)\) as random
Measuring Robustness (2/3)

- We repeated the experiment of [Santus et al., 2016a]
  - Adding switched hypernym pairs as random
  - e.g. \((apple, animal), (cow, fruit)\)

- Using BLESS (the only dataset with random pairs)
- For each hypernym pair \((x_1, y_1)\):
  - Sampled \(y_2\) such that \((x_2, y_2) \in\) dataset is a hypernym-pair
  - such that \((x_1, y_2) \notin\) dataset
  - Add \((x_1, y_2)\) as random

- We added 139 such new random pairs
Measuring Robustness (2/3)

- We repeated the experiment of [Santus et al., 2016a]
  - Adding switched hypernym pairs as random
  - e.g. \((\text{apple, animal}), (\text{cow, fruit})\)

- Using BLESS (the only dataset with random pairs)
  - For each hypernym pair \((x_1, y_1)\):
    - Sampled \(y_2\) such that \((x_2, y_2) \in\) dataset is a hypernym-pair
    - such that \((x_1, y_2) \notin\) dataset
    - Add \((x_1, y_2)\) as random

- We added 139 such new random pairs
- We used the best supervised and unsupervised methods to re-classify the revised dataset
Supervised method drops in performance, unsupervised performance is stable:

<table>
<thead>
<tr>
<th>method</th>
<th>AP@100 original</th>
<th>AP@100 switched</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>concat, word2vec, $L_1$</td>
<td>0.995</td>
<td>0.575</td>
</tr>
<tr>
<td>unsupervised</td>
<td>cosWeeds, win2d, ppmi</td>
<td>0.818</td>
<td>0.882</td>
</tr>
</tbody>
</table>
Measuring Robustness (3/3)

- Supervised method drops in performance, unsupervised performance is stable:

<table>
<thead>
<tr>
<th>method</th>
<th>AP@100 original</th>
<th>AP@100 switched</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>0.995</td>
<td>0.575</td>
<td>-0.42</td>
</tr>
<tr>
<td>unsupervised</td>
<td>0.818</td>
<td>0.882</td>
<td>+0.064</td>
</tr>
</tbody>
</table>

- 121 out of the 139 switched pairs were falsely classified as hypernyms by the supervised method!
Supervised method drops in performance, unsupervised performance is stable:

<table>
<thead>
<tr>
<th>method</th>
<th>AP@100 original</th>
<th>AP@100 switched</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>concat, word2vec, $L_1$</td>
<td>0.995</td>
<td>0.575</td>
<td>-0.42</td>
</tr>
<tr>
<td>unsupervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cosWeeds, win2d, ppmi</td>
<td>0.818</td>
<td>0.882</td>
<td>+0.064</td>
</tr>
</tbody>
</table>

121 out of the 139 switched pairs were falsely classified as hypernyms by the supervised method!

Unsupervised measures are robust, performance even slightly improves
Measuring Robustness (3/3)

- Supervised method drops in performance, unsupervised performance is stable:

<table>
<thead>
<tr>
<th>method</th>
<th>AP@100 original</th>
<th>AP@100 switched</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>concat, word2vec, $L_1$</td>
<td>0.995</td>
<td>0.575</td>
</tr>
<tr>
<td>unsupervised</td>
<td>cosWeeds, win2d, ppmi</td>
<td>0.818</td>
<td>0.882</td>
</tr>
</tbody>
</table>

- 121 out of the 139 switched pairs were falsely classified as hypernyms by the supervised method!
- Unsupervised measures are robust, performance even slightly improves
- Consistent with the transfer-learning experiment of [Turney and Mohammad, 2015]
Recap

- We experimented with an extensive number of unsupervised measures for hypernymy detection.
Recap

- We experimented with an extensive number of unsupervised measures for hypernymy detection
- There is no specific measure which is always preferred
- Unsupervised measures are not deprecated given supervised methods
- Supervised methods are favorable by the existing datasets
- However, they are extremely sensitive to training data
- Unsupervised methods are more robust
Recap

- We experimented with an extensive number of unsupervised measures for hypernymy detection

- There is no specific measure which is always preferred
  - Different measures capture different aspects of hypernymy
Recap

- We experimented with an extensive number of unsupervised measures for hypernymy detection.

- There is no specific measure which is always preferred.
  - Different measures capture different aspects of hypernymy.
  - Specific measures are preferred in distinguishing hypernymy from each other relation.

- Unsupervised measures are not deprecated given supervised methods.
- Supervised methods are favorable by the existing datasets.
- However, they are extremely sensitive to training data.
- Unsupervised methods are more robust.
Recap

- We experimented with an extensive number of unsupervised measures for hypernymy detection

- There is no specific measure which is always preferred
  - Different measures capture different aspects of hypernymy
  - Specific measures are preferred in distinguishing hypernymy from each other relation

- Unsupervised measures are not deprecated given supervised methods
Recap

- We experimented with an extensive number of unsupervised measures for hypernymy detection.

- There is no specific measure which is always preferred.
  - Different measures capture different aspects of hypernymy.
  - Specific measures are preferred in distinguishing hypernymy from each other relation.

- Unsupervised measures are not deprecated given supervised methods.
  - Supervised methods are favorable by the existing datasets.
Recap

▶ We experimented with an extensive number of unsupervised measures for hypernymy detection

▶ There is no specific measure which is always preferred
  ▶ Different measures capture different aspects of hypernymy
  ▶ Specific measures are preferred in distinguishing hypernymy from each other relation

▶ Unsupervised measures are not deprecated given supervised methods
  ▶ Supervised methods are favorable by the existing datasets
  ▶ However, they are extremely sensitive to training data
We experimented with an extensive number of unsupervised measures for hypernymy detection.

There is no specific measure which is always preferred.
- Different measures capture different aspects of hypernymy.
- Specific measures are preferred in distinguishing hypernymy from each other relation.

Unsupervised measures are not deprecated given supervised methods.
- Supervised methods are favorable by the existing datasets.
- However, they are extremely sensitive to training data.
- Unsupervised methods are more robust.
Future Work

- A combination of unsupervised measures may perform well in distinguishing hypernym from all other relations.
Future Work

- A combination of unsupervised measures may perform well in distinguishing hypernym from all other relations
- A simple classifier with unsupervised measures as features performs well
Future Work

▶ A combination of unsupervised measures may perform well in distinguishing hypernym from all other relations
▶ A simple classifier with unsupervised measures as features performs well
▶ However, it’s not as robust as the unsupervised measures
Future Work

- A combination of unsupervised measures may perform well in distinguishing hypernym from all other relations.
- A simple classifier with unsupervised measures as features performs well.
- However, it’s not as robust as the unsupervised measures.
- Current datasets are biased and encourage overfitting.

Thank you!
Future Work

- A combination of unsupervised measures may perform well in distinguishing hypernym from all other relations
- A simple classifier with unsupervised measures as features performs well
- However, it's not as robust as the unsupervised measures
- Current datasets are biased and encourage overfitting
- In the future, the development of natural datasets may better reflect the performance of methods in real-application scenarios

Thank you!
Future Work

▶ A combination of unsupervised measures may perform well in distinguishing hypernym from all other relations
▶ A simple classifier with unsupervised measures as features performs well
▶ However, it’s not as robust as the unsupervised measures
▶ Current datasets are biased and encourage overfitting
▶ In the future, the development of natural datasets may better reflect the performance of methods in real-application scenarios

Thank you!
References


