An Evaluation of Graded Sense Disambiguation using Word Sense Induction

David Jurgens
HRL Laboratories, LLC and
University of California, Los Angeles
Talk overview

• Word Sense Disambiguation and Graded Senses Assignment
• Word Sense Induction
• Evaluating Graded Sense Assignments
Talk overview

- Word Sense Disambiguation and Graded Senses Assignment
- Word Sense Induction
- Evaluating Graded Sense Assignments
Word Sense Disambiguation

**Goal:** Given a usage of a word, determine which meaning is present

John sat on the **chair**.

1. a seat for one person, with a support for the back
2. the position of professor
3. the officer who presides at the meetings of an organization
Polysemy makes sense disambiguation much harder.

Mary handed the **paper** to her professor.

1. a material made of cellulose pulp
2. an essay, especially one written as an assignment
3. a daily or weekly publication on folded sheets
4. a medium for written communication
5. a scholarly article describing the results of observations
6. a business firm that publishes newspapers
Polysemy makes sense disambiguation much harder.

Mary handed the paper to her professor.

1. a material made of cellulose pulp
2. an essay, especially one written as an assignment
3. a daily or weekly publication on folded sheets
4. a medium for written communication
5. a scholarly article describing the results of observations
6. a business firm that publishes newspapers
Polysemy makes sense disambiguation much harder.

After working all night writing her assignment, Mary handed the paper to her professor.

1. a material made of cellulose pulp
2. an essay, especially one written as an assignment
3. a daily or weekly publication on folded sheets
4. a medium for written communication
5. a scholarly article describing the results of observations
6. a business firm that publishes newspapers
Polysemy makes sense disambiguation much harder.

After working all night writing her assignment, Mary handed the paper to her professor.

1. a material made of cellulose pulp
2. an essay, especially one written as an assignment
3. a daily or weekly publication on folded sheets
4. a medium for written communication
5. a scholarly article describing the results of observations
6. a business firm that publishes newspapers

Most applicable sense

Moderately applicable sense
The goal of Graded Word Sense Disambiguation

- Identify which senses are salient in a given context
- Quality the senses’ degree of applicability
Talk overview

• Word Sense Disambiguation and Graded Senses Assignment

• Word Sense Induction

• Evaluating Graded Sense Assignments
Word Sense Disambiguation

- Sense Inventory
- New Context
- Sense Classifier
- Context
- Sense Annotated Corpus
- Annotation
Word Sense Induction automatically learns the sense inventory from contextual examples

**Sense Induction**

- **Unlabeled Corpus**
- **Unsupervised Learning**
- **Sense Inventory**
- **Sense Classifier**
- **New Context**
- **Sense Annotated Corpus**
- **Annotation**
Benefits of Sense Induction

• Discovery of corpus-specific senses

• Removes the need for learning sense features from sense-annotated corpora

• No more annotation bottleneck!
Potential Drawbacks

• What senses are being learned?

• How useful are the learned senses?
Induced senses can be used directly, but often a specific sense inventory is needed.
Induced senses can be used directly, but often a specific sense inventory is needed.

Induced Sense Inventory with features

Graded Word Sense Disambiguation

Graded Annotations with Induced Senses

Sense Mapping Function

Contexts Labeled with Induced and Reference senses

Example Context

Produces an end-to-end WSD system with the desired sense labels, using minimal labeled data.
Three graph-based WSI methods tested

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns, Collocated Nouns</td>
<td>Nouns, Verbs, and Adjectives</td>
<td>Nouns, Verbs, and Adjectives</td>
<td></td>
</tr>
<tr>
<td>Sense Discovery Method</td>
<td>Single Word</td>
<td>All Words</td>
<td>Single Word</td>
</tr>
<tr>
<td>Chinese Whispers (Biemann, 2006)</td>
<td>Link Clustering (Ahn, 2010)</td>
<td>Edge Deletion</td>
<td></td>
</tr>
</tbody>
</table>
Sense Remapping at a high level

1. Co-label corpus with both sense types
2. Build a collocation matrix for senses
3. Construct a function to remap specific instances

\[
i_i \rightarrow \{r_j\}
\]
Naïve approach: map each induced sense to a single sense

**Step 1:** Count co-labeling of induced and reference senses

<table>
<thead>
<tr>
<th></th>
<th>r₁</th>
<th>r₂</th>
<th>r₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>i₁</td>
<td>6</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>i₂</td>
<td>10</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>i₃</td>
<td>2</td>
<td>16</td>
<td>2</td>
</tr>
</tbody>
</table>

**Step 2:** Normalize row values to be \( p(r_i|i_j) \).

<table>
<thead>
<tr>
<th></th>
<th>r₁</th>
<th>r₂</th>
<th>r₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>i₁</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>i₂</td>
<td>0.625</td>
<td>0.125</td>
<td>0.25</td>
</tr>
<tr>
<td>i₃</td>
<td>0.1</td>
<td>0.8</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Naïve approach: map each induced sense to a single sense

<table>
<thead>
<tr>
<th></th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_1$</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>$i_2$</td>
<td>0.625</td>
<td>0.125</td>
<td>0.25</td>
</tr>
<tr>
<td>$i_3$</td>
<td>0.1</td>
<td>0.8</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Step 3: Map each induced sense to the most likely reference sense

$i_1 \rightarrow r_2$

$i_2 \rightarrow r_1$

$i_3 \rightarrow r_2$
Naïve approach: map each induced sense to a single sense

Context with Graded Sense Labels

\[
\begin{array}{ccc}
  i_1 & i_2 & i_3 \\
  0.7 & 0.2 & 0.1 \\
\end{array}
\]

Sum the ratings of each of the remapped senses

\[
\begin{array}{ccc}
  r_1 & r_2 & r_3 \\
  0.2 & 0.8 & 0 \\
\end{array}
\]
Sense mapping from Agirre et al. (2006)

**Step 3:** multiply graded labeling by mapping matrix

\[
\begin{array}{ccc}
  i_1 & i_2 & i_3 \\
  0.3 & 0.5 & 0.2 \\
\end{array}
\]

\[
\begin{array}{ccc}
  r_1 & r_2 & r_3 \\
  0.3 & 0.6 & 0.1 \\
  0.625 & 0.125 & 0.25 \\
  0.1 & 0.1 & 0.8 \\
\end{array}
\]

\[
\begin{array}{ccc}
  r_1 & r_2 & r_3 \\
  0.423 & 0.263 & 0.315 \\
\end{array}
\]

Select the highest weighted sense
**Step 3**: multiply graded labeling by mapping matrix

<table>
<thead>
<tr>
<th></th>
<th>(i_1)</th>
<th>(i_2)</th>
<th>(i_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r_1)</td>
<td>0.3</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>(r_2)</td>
<td>0.625</td>
<td>0.125</td>
<td>0.25</td>
</tr>
<tr>
<td>(r_3)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

\[ \begin{array}{ccc}
0.3 & 0.5 & 0.2 \\
0.625 & 0.125 & 0.25 \\
0.1 & 0.1 & 0.8 \\
\end{array} \]  

\[ \begin{array}{ccc}
0.423 & 0.263 & 0.315 \\
\end{array} \]

Use all the senses
**Issue:** the matrix construction method ignores the graded ratings

<table>
<thead>
<tr>
<th>Contexts</th>
<th>i₁ : 0.9</th>
<th>i₂ : 0.1</th>
<th>r₁ : 0.9</th>
<th>r₂ : 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>i₁ : 0.9</td>
<td>i₂ : 0.1</td>
<td>r₁ : 0.9</td>
<td>r₂ : 0.1</td>
</tr>
<tr>
<td>C₂</td>
<td>i₁ : 0.1</td>
<td>i₂ : 0.9</td>
<td>r₁ : 0.1</td>
<td>r₂ : 0.9</td>
</tr>
<tr>
<td>C₃</td>
<td>i₁ : 0.1</td>
<td>i₂ : 0.9</td>
<td>r₁ : 0.1</td>
<td>r₂ : 0.9</td>
</tr>
</tbody>
</table>

Matrix:

<table>
<thead>
<tr>
<th></th>
<th>r₁</th>
<th>r₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>i₁</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>i₂</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
**Alternative:** construct the co-label matrix by summing the product of the sense ratings

<table>
<thead>
<tr>
<th>Contexts</th>
<th>(i_1: 0.9)</th>
<th>(r_1: 0.9)</th>
<th>(i_2: 0.1)</th>
<th>(r_2: 0.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(i_1: 0.1)</td>
<td>(r_1: 0.1)</td>
<td>(i_2: 0.9)</td>
<td>(r_2: 0.9)</td>
</tr>
<tr>
<td>(C_2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(i_1: 0.1)</td>
<td>(r_1: 0.1)</td>
<td>(i_2: 0.9)</td>
<td>(r_2: 0.9)</td>
</tr>
<tr>
<td>(C_3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(i_1: 0.1)</td>
<td>(r_1: 0.1)</td>
<td>(i_2: 0.9)</td>
<td>(r_2: 0.9)</td>
</tr>
</tbody>
</table>

Compute the subsequent steps the same as Agirre et al. (2006)
Equivalent to the Naïve approach with a vector mapping

**Steps 1 and 2:** Construct row-normalized matrix as before

<table>
<thead>
<tr>
<th></th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_1$</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>$i_2$</td>
<td>0.625</td>
<td>0.125</td>
<td>0.25</td>
</tr>
<tr>
<td>$i_3$</td>
<td>0.1</td>
<td>0.8</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**Step 3:** Map each induced sense to a distribution over senses

<table>
<thead>
<tr>
<th></th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_1$</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>$i_2$</td>
<td>0.625</td>
<td>0.125</td>
<td>0.25</td>
</tr>
<tr>
<td>$i_3$</td>
<td>0.1</td>
<td>0.8</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Equivalent to the Naïve approach with a vector mapping

Context with Graded Sense Labels

<table>
<thead>
<tr>
<th>i_1</th>
<th>i_2</th>
<th>i_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Sum the induced senses’ distributions, weighted by their ratings

\[
\begin{align*}
\text{Context with Graded Sense Labels} & \\
\begin{array}{ccc}
  r_1 & r_2 & r_3 \\
  0.3 & 0.6 & 0.1 \\
  0.625 & 0.125 & 0.25 \\
  0.1 & 0.8 & 0.1 \\
\end{array}
\end{align*}
\]

\[
\begin{align*}
\text{Sum the induced senses’ distributions, weighted by their ratings} & \\
\begin{array}{ccc}
  r_1 & r_2 & r_3 \\
  0.345 & 0.525 & 0.13 \\
\end{array}
\end{align*}
\]
Remapping Summary

Three Approaches

- Naïve
- Agirre et al. (2006)
- Vector
- Vector + Matrix reweighting

![Diagram showing three approaches: Naïve, Agirre et al. (2006), Vector, Vector + Matrix reweighting]
Talk overview

• Word Sense Disambiguation and Graded Senses Assignment
• Word Sense Induction
• Evaluating Graded Sense Assignments
Evaluating Graded Sense Annotation

She *won* the gold medal with her hard work.

\[
\begin{align*}
\text{wn}_1 &: 0.6 \\
\text{wn}_2 &: 0.4 \\
\text{wn}_3 &: 0.0 \\
\text{wn}_4 &: 0.0
\end{align*}
\]
Multiple evaluation objectives

She *won* the gold medal with her hard work.

1. Which senses are applicable?

| wn₁: 0.6 |
| wn₂: 0.4 |
| wn₃: 0.0 |
| wn₄: 0.0 |
She \textit{won} the gold medal with her hard work.

\begin{itemize}
\item \textit{wn}_1: 0.6
\item \textit{wn}_2: 0.4
\item \textit{wn}_3: 0.0
\item \textit{wn}_4: 0.0
\end{itemize}

1. Which senses are applicable?
2. How do senses differ in their applicability?
Multiple evaluation objectives

She *won* the gold medal with her hard work.

1. Which senses are applicable?
2. How do senses differ in their applicability?
3. What is the applicability of each sense?

| wn1: 0.6   | 1. Which senses are applicable? |
| wn2: 0.4   | 2. How do senses differ in their applicability? |
| wn3: 0.0   | 3. What is the applicability of each sense? |
| wn4: 0.0   |                                      |
Detection: which senses are present?

She \textit{won} the gold medal with her hard work.

<table>
<thead>
<tr>
<th></th>
<th>Gold</th>
<th>Test$_1$</th>
<th>Test$_2$</th>
<th>Test$_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>wn$_1$:</td>
<td>0.6</td>
<td>0.7</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>wn$_2$:</td>
<td>0.4</td>
<td>0.3</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>wn$_3$:</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>wn$_4$:</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Jaccard Index

\[
\frac{G \cap T}{G \cup T}
\]

Detection: 1.0 0.5 0.25
Ranking: order the senses by their applicability

She won the gold medal with her hard work.

<table>
<thead>
<tr>
<th></th>
<th>Gold</th>
<th>Test(_1)</th>
<th>Test(_2)</th>
<th>Test(_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wn(_1):</td>
<td>0.6</td>
<td>0.7</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>wn(_2):</td>
<td>0.4</td>
<td>0.3</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>wn(_3):</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>wn(_4):</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Goodman and Kruskal’s \(\gamma\)

- Computed much like Kendall’s \(\tau\)
- Most appropriate for comparing rankings with many ties

Ranking: 1.0  0.33  -0.2
Perception: quantify the applicability of each sense

She *won* the gold medal with her hard work.

<table>
<thead>
<tr>
<th>Gold</th>
<th>Test₁</th>
<th>Test₂</th>
<th>Test₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>wn₁: 0.6</td>
<td>wn₁: 0.7</td>
<td>wn₁: 0.0</td>
<td>wn₁: 0.3</td>
</tr>
<tr>
<td>wn₂: 0.4</td>
<td>wn₂: 0.3</td>
<td>wn₂: 1.0</td>
<td>wn₂: 0.0</td>
</tr>
<tr>
<td>wn₃: 0.0</td>
<td>wn₃: 0.0</td>
<td>wn₃: 0.0</td>
<td>wn₃: 0.5</td>
</tr>
<tr>
<td>wn₄: 0.0</td>
<td>wn₄: 0.0</td>
<td>wn₄: 0.0</td>
<td>wn₄: 0.2</td>
</tr>
</tbody>
</table>

Cosine Similarity:

\[
\frac{G \cdot T}{||G|| \cdot ||T||}
\]

Perception: 0.98 0.55 0.41
Baselines

Sense assignments have a strong bias towards more frequent senses

- Most Frequent Sense (MFS)
- All Senses, ranked by Frequency (ASF)
- All Senses, equally ranked (ASE)
Experimental Setup

- Trained using the GWS Corpus provided by Erk et al. (2009)
- 8 terms: 3 nouns, 3 verbs, 2 adj. (4-7 senses each)
- 50 contexts each
- 3 annotators per context, average sense scores used as gold standard
- Used 80% to build sense mapping, 20% to test with cross validation
Perception Results

- MFS
- ASE
- ASF
- Jurgens (2011)
- Navigli and Crisafulli (2010)
- Korkontzelos and Manandhar (2010)
Ranking Results

- MFS
- ASE
- ASF
- Jurgens (2011)
- Navigli and Crisafulli (2010)
- Korkontzelos and Manandhar (2010)
Perception Results

- MFS
- ASE
- ASF
- Jurgens (2011)
- Navigli and Crisafulli (2010)
- Korkontzelos and Manandhar (2010)
Further Observations

• Approaches varied wildly in the number of senses labeled
  • Navigli and Crisafulli (2010) only labeled ~56%, but had very high precision (only 5% had > 1 label)
  • Jurgens (2011) ~2 senses on average to four words, and 60+ for the other four

• Ranking is much harder than Perception
  • A > B vs. A=.51 and B=.49
Take-aways

• WSI systems offer significant promise for automating Graded WSD

• Seeding corpus annotation when WSD systems have too little training data

• Accounting for graded rating is essential in sense-mapping
Future Work

• Rethinking the *Perception* evaluation metric
  • Neither Cosine similarity nor Jensen-Shannon Diverge appear to be ideal measures

• Intrinsic clustering evaluations for partial cluster membership

• Task-based Evaluation for graded senses
Thank you!

jurgens@cs.ucla.edu

All models and data released as a part of the S-Space Package

https://github.com/fozziethebeat/S-Space

Thanks to Katrin Erk, Diana McCarthy, and Nicholas Gaylord for making GWS corpus available