LSCDiscovery: A shared task on semantic change discovery and detection in Spanish

Frank D. Zamora-Reina ¹  Felipe Bravo-Marquez ¹  Dominik Schlechtweg ²

¹Department of Computer Science, University of Chile, IMFD & CENIA

²Institute for Natural Language Processing, University of Stuttgart

May 2022
LSCDiscovery: Phases

1. Graded Change Discovery,

2. Binary Change Detection
Given a diachronic corpus pair $C_1$ and $C_2$, rank the intersection of their (content-word) vocabularies according to their degree of change between $C_1$ and $C_2$. 
Given a target word $w$ and two sets of its usages $U_1$ and $U_2$, decide whether $w$ lost or gained senses from $U_1$ to $U_2$, or not.
• *discovery* introduces additional difficulties for models
  • a large number of predictions is required
  • target word are not preselected, balanced or cleaned
LSCDiscovery: Optional Tasks

- Graded Change Detection
- Sense Gain Detection
- Loss Gain Detection
- COMPARE
Graded Change Detection

- similar to Graded Discovery
- the only difference was the public target words corresponded exactly to the hidden words on which we evaluated
Sense Gain Detection

- similar to Binary Change Detection
- only words which gained (not lost) senses receive label 1.
Sense Loss Detection

- similar to Binary Change
- only words which lost (not gained) senses received label 1.
LSCDiscovery: Optional Tasks

COMPARE

- average of human semantic proximity judgments of usage pairs
- approximation of JSD (Graded Change)
# LSCDiscovery: Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Time period</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old corpus (C1)</td>
<td>1810–1906</td>
<td>(\sim 13M)</td>
</tr>
<tr>
<td>Modern corpus (C2)</td>
<td>1994–2020</td>
<td>(\sim 22M)</td>
</tr>
</tbody>
</table>

**Table:** Sizes of both corpora.
LSCDiscovery: Previous Shared Tasks

- SemEval 2020 Task 1
- DIACR-Ita
- RuShiftEval
### LSCDiscovery: Previous Shared Tasks

<table>
<thead>
<tr>
<th>Shared Task</th>
<th>Target words (dev/testing)</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemEval 2020 Task 1</td>
<td>0/156</td>
<td>Binary Change/Graded Change</td>
</tr>
<tr>
<td>DIACR-Ita</td>
<td>0/18</td>
<td>Binary Change Detection</td>
</tr>
<tr>
<td>RuShiftEval</td>
<td>12/99</td>
<td>COMPARE</td>
</tr>
</tbody>
</table>
public target words → 4385 (only evaluation phase 1)
hidden target words → 80
  development set → 20
  evaluation set → 60
exact annotated target word usages were provided
12 annotators

each target word was sampled \( |U_1| = |U_2| = 20 \) usages (sentences) per subcorpus \((C_1, C_2)\).

\( \sim 62K \) judgments

- 12k judgments for development
- 38k judgments for evaluation
- 12k judgments for discarded (due to the low agreement)
Figure: Word Usage Graph of Spanish *servidor*.
Graded Change Discovery

- Spearman correlation
COMPARE

- Spearman correlation
Binary Change Detection

- F1 (main metric)
- Precision
- Recall
Sense Gain Detection

- F1
- Precision
- Recall
LSCDiscovery: Metrics

Sense Loss Detection

- F1
- Precision
- Recall
**Baselines**

- **baseline1**: SGNS + OP + CD
- **baseline2**: Normalized Log-Transformed Frequency Difference
- **baseline3**: Grammatical Profiling
- **baseline4**: Minority class
- **baseline5**: Random baseline
GlossReader (token-based system)
- fine-tuned the XLM-R multilingual as part of a gloss-based Word Sense Disambiguation (WSD) system language model

DeepMistake (token-based system)
- WiC model, initially trained by fine-tuning the XLM-R model

HSE (token-based system)
- fine-tuning BERT, and then clustering using K-means
LSCDiscovery: Winning systems in phase 2

- GlossReader (token-based system)
- UAAlberta (token-based and type-based system)
  - SGNS + XLM-R + APD
- Rombek (token-based system)
  - WSI task
the winning system for phase 1 and 2 actually models the COMPARE score with APD
for phase 2 it uses thresholding on the graded scores
• performance for **graded change** comparable to previous shared tasks
  • but obtained under harder conditions (Discovery)
    → applicable to solve real-world historical semantics/lexicography problems

• performance for **binary change** lower, but still above baseline
  • more relevant to historical semantics/lexicography
    → future challenge

• both tasks dominated by token-based models
  → confirms tendency observed in RuShiftEval

• clustering methods amongst the best-performing systems for the first time
  • important, because current systems exploit correlations between change measures and do not model annotation procedure
    → **upper performance bound**
END