### Introduction

Crowdsourced linguistic data is often noisy in some way. Instead of throwing out unsuitable data, we show that adaptation of NLP tools to the data allows automatic processing with good accuracy for a narrow but unknown domain, keeping more information. This poster demonstrates the concept for spell-checking and pronoun resolution tools.

### Event Sequence Descriptions

Our data is a web-collected data set of descriptions of kitchen tasks, gathered for script mining [2]. A script describes such a scenario:

1. First strip of the papery skin of the bulb
2. Ease out as many intact cloves as possible
3. Chop them finely if you want a stronger taste
4. Chop them coarsely if you want a weaker taste
5. Crush garlic is the strongest taste

Example of one web-collected kitchen task description:

1. First strip of the papery skin of the bulb
2. Ease out as many intact cloves as possible
3. Chop them finely if you want a stronger taste
4. Chop them coarsely if you want a weaker taste
5. Crush garlic is the strongest taste

### Script Mining Task

- Many web-collected descriptions are used to create a single model script
- Different descriptions should be matched:
  - chop them = chip the garlic up.
- This requires preprocessing:
  - Spelling correction: Makes for better input for standard applications or dictionaries.
  - Pronoun resolution: To know what the object in 'chop them' is.

### Pronoun Resolution

To adapt the spelling correction tool to our data, we use an unmodified general tool but add domain-sensitive heuristics to select detections and corrections.

**General use:** GNU Aspell

- Uses a general dictionary (which may lack kitchen-specific terms or names)
- This leads to false corrections of correct words such as:
  - deglaze -> dry lase
  - microplane grater -> micro plane grater
  - siplot -> zoology

**Modified Aspell**

We also consider the domain context (slice bread):

- If a word occurs in at least 3 other descriptions, the system accepts it
- Only accept corrections that occur in at least 1 other description, preventing:
  - bread -> bird (bird), leaflet -> loft (leaf)
- Split only if the resulting words occur in another description, preventing:
  - sandwich -> sand which sandwich

The intended word is in parentheses. These off-topic corrections would be plausible in different domains than ‘slice bread’.

### Evaluation

Evaluation based on manual judgement of the corrections made by the spell-checkers.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>False Positives</th>
<th>Type Precision</th>
<th>Sem. Precision</th>
<th>Corrections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspell</td>
<td>0.43</td>
<td>0.28</td>
<td>0.57</td>
<td>0.58</td>
<td>162</td>
</tr>
<tr>
<td>Enhanced Aspell</td>
<td>0.58</td>
<td>0.29</td>
<td>0.79</td>
<td>0.76</td>
<td>150</td>
</tr>
</tbody>
</table>

### References