Inconsistency Detection in Job Postings

Joana Urbano, skeeped, Luxembourg; Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Faculdade de Engenharia, Universidade do Porto, Portugal
Miguel Couto, skeeped, Luxembourg
Gil Rocha, Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Faculdade de Engenharia, Universidade do Porto, Portugal
Henrique Lopes Cardoso, Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Faculdade de Engenharia, Universidade do Porto, Portugal

Overview
- The use of AI in recruitment is growing;
- AI software can read jobs' descriptions and select the best candidates for these jobs;
- These descriptions may be ambiguous and/or contain contradictions between unstructured and structured fields.

Contributions:
- A terminology for inconsistencies in the description of language requirements in English job postings.
- A model based on NLP, machine learning and rule-based approaches.

Example
Unstructured Input: “The candidate must have a masters and experience in biology, biochemistry or related areas. We expect good knowledge of English and similar knowledge of either French or Portuguese. German is considered an asset.”

Structured input:

<table>
<thead>
<tr>
<th>language</th>
<th>level</th>
<th>optional</th>
<th>alternative</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>B2</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>fr</td>
<td>B2</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>pt</td>
<td>B2</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>de</td>
<td>B1</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

When comparing the structured and unstructured inputs, we can obtain several types of inconsistencies:

Language-related inconsistencies:
- Language-not-specified contradiction
- Language-not-required contradiction
- Language-not-optional contradiction
- Lexical contraction
- Numerical contraction
- Alternative-language contradiction
- Ambiguity

Methodology

1. Sentence segmentation and selection
   - Preprocessing of textual descriptions
   - Sentence segmentation
   - Language-related sentence selection

2. Language extraction from textual descriptions
   - Language disambiguation
   - IT for Italian or information technology?
   - Our solution: use of controlled vocabulary to answer these ambiguities
   - Random forest model:
     - one-hot encoding
     - language-related features:
       - language names;
       - language modifiers (e.g. “fluent”);
       - others (e.g. “speak”, “write”).
   - controlled vocabulary for:
     - possible language modifiers;
     - optional related words (e.g. “asset”, “plus”);
     - non-optional related words (e.g. “mandatory”);
     - alternative related words (e.g. “either... or”);
     - identification of patterns
     - rule-based approach

3. Language extraction from structured fields
   - Straightforward language-related information extraction
   - No validation needed/required

4. Inconsistency detection
   - Modifiers to language levels conversion
   - Matching between textual descriptions and structured fields

Conclusion
- Proposed a terminology for the description of inconsistencies in language requirements;
- Proposed a 4-step NLP-based model to detect them in job descriptions, combining both machine learning and rule-based approaches.
- The model achieved high performance on each step.

Future work:
- Replace rule-based approach with ML;
- Extend our annotated dataset of job postings;
- Adapt the model to text written in other languages.

Results - step 1
Train/test data description:

<table>
<thead>
<tr>
<th>sentences</th>
<th>jobs</th>
<th>positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>437</td>
<td>478 499</td>
</tr>
<tr>
<td>test</td>
<td>892</td>
<td>88  84</td>
</tr>
</tbody>
</table>

Performance on test data:
- accuracy: 99.21%
- recall: 95.24%
- F1-score: 95.81%

Results - step 2
Train/test data description:

Performance on test data:

<table>
<thead>
<tr>
<th>label</th>
<th>errors</th>
<th>accuracy</th>
<th>recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>modifier</td>
<td>0.00%</td>
<td>99.00%</td>
<td>99.27%</td>
<td>99.21%</td>
</tr>
<tr>
<td>required languages</td>
<td>0.00%</td>
<td>99.00%</td>
<td>99.27%</td>
<td>99.21%</td>
</tr>
</tbody>
</table>

Results - step 4
Train/test data description:

Performance on test data:
- accuracy: 100.00%

Acknowledgements
This research is partially supported by LIACC (FCT/UID/CEC/0027/2020) funded by Fundação para a Ciência e a Tecnologia (FCT). Gil Rocha is supported by a PhD grant (with reference SFRH/BD/140125/2018) from FCT. We thank Gálica Correia for her contribution in the initial phase of this project.