Inconsistency Detection in Job Postings

Joana Urbano, skeeled, Luxembourg; Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Faculdade de Engenharia, Universidade do Porto, Portugal
Miguel Couto, skeeled, 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 domain specific rules to detect these inconsistencies.

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:

language	level	optional	alternative
en	B2	no	no
fr	B2	no	yes
pt	B2	no	yes
de	B1	yes	–
Required languages: 2			

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 contradiction
- Numerical contradiction
- Alernative-language contradiction
- Ambiguity

Methodology

1. Sentence segmentation and selection

Preprocessing of textual descriptions
↓
Sentence segmentation
↓
Language-related sentence selection

<u>Random forest model</u>:
- one-hot encoding
- language-related features:
  - language names;
  - language modifiers (e.g. "fluent");
  - others (e.g. "speak", "write").

2. Language extraction from textual descriptions

Language disambiguation

IT for Italian or information technology?

Our solution: use of controlled vocabulary to answers these ambiguities

↓
Language-related information extraction

- 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:

	sentences	jobs	positive
train	4267	478	490
test	882	88	84

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

Results - step 2

Train/test data description:

	entries	sentences	jobs
train	529	262	216
test	204	109	86

Performance on test data:

label	#errors	accuracy
modifier	fluent: 4, fluency: 4, good: 1, knowledge: 1, others: 0	94.90%
required languages	0: –, 1: 0, 2: 1, 3: 1, 4: 0	97.67%

	accuracy	recall	f1-score
Optional/Non-Optional	94.9%	98.04%	96.77%
Alternative/Non-Alternative	98.09%	100%	98.96%
Ambiguity	94.18%	83.33%	66.67%

Results - step 4

Train/test data description:

	entries	sentences	jobs
train	519	252	216
test	196	101	86

Performance on test data:
- accuracy: 100% for
  - Language-not-specified,
  - Language-not-required,
  - Language-not-optional,
  - Alternative language
  - Numerical

Lexical inconsistency:
- accuracy: 98.08%
- recall: 91.30%
- F1-score: 95.45%.

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 Catarina Correia for her contribution in the initial phase of this project.