Combining Grammatical and Relational Approaches. A Hybrid Method for the Identification of Candidate Collocations from Corpora

Damiano Perrin, Irene Fioravanti, Osvaldo Gervasi, Stefania Spina -University of Perugia, -University for Foreigners of Perugia, Italy
damiano.perrin, osvaldo.gervasi@unipg.it
irene.fioravanti, stefania.spina@unistraspg.it

The research has been funded by the Italian Ministry of Research (PRIN) PRIN Research Projects of Major National Interest - Call 2022 - Pres.2022X5AE
The title of the project is DICI-A: A Learner Dictionary of Italian Collocations

Background

NLP techniques are a powerful tool for identifying candidate collocations in corpora for the development of lexicographic resources (Evert, 2004).

Two main methods:

• P-based approach
  • Reliance on Part-Of-Speech tagging.
  • Improvement in detection accuracy with POS filter (Krieny, 2000; Ritz, 2006).
  • Failure in detecting non-adjacent word pairs (Seretan, 2011).

• S-based approach
  • Utilisation of syntactic dependencies for capturing discontinuous collocations.
  • Challenges with parsing accuracy affecting detection (Lu & Zhou, 2004).

Call for hybrid approaches: combining P-based and S-based methods for incrementing detection accuracy (Castagnoli et al., 2016).

Method

Two types of collocations: Vdobj (verb + direct object) and Vmod (adjective modifier).

Sample texts
Eight texts randomly extracted from the Perugia corpus (Spina, 2014) of a total of ca. 8000 tokens balanced across registers and text genres.

Three systems
• P-based approach: texts were pos-tagged with Tree Tagger and searched with the Corpus Workbench tool and the Corpus Query Processing > 549 candidates.
• S-based approach: texts were parsed with the spaCy library > 685 candidates.
• Hybrid approach: merge of the two previous methods > 748 candidates.

The benchmark was obtained through a human annotation process > 610 candidates.

The Main Research Question

Presenting a hybrid approach to detecting candidate collocations from corpora for the development of a learner dictionary of Italian collocations.

Does the hybrid approach perform better in the candidate identification task compared to the P-based and the S-based approach?

Computational Procedure

Two steps

• Pre-processing of the input text for the standardisation of the input data format to remove any irrelevant elements.
• Sentence parsing with spaCy and implementation of rules to optimise analyses. For example, the following function is designed to identify AMOD when the Vmod relation exists, with 'obj' as the dependency, and the UPosTag of the 'obj' token in NOUN:

  if token.dep == "amod" and
  token.head.dep == "obj" and
  token.pos == "ADJ" and
  token.head.pos == "NOUN"

Analyses

• Evaluation of the three approaches compared through measures of accuracy, precision, recall and F1 score.
• Computation of the benchmark match to estimate how well the model aligns with the correct prediction established by the benchmark annotation:

  Bm = 100 * (TP+TN) / (TP+TN+FN)

  [TP=true positive; TN = true negative; FN= false negative]

Table 1. Comparison of the three methods concerning Vdobj

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-based</td>
<td>0.76</td>
<td>0.83</td>
<td>0.90</td>
<td>0.87</td>
</tr>
<tr>
<td>S-based</td>
<td>0.68</td>
<td>0.88</td>
<td>0.75</td>
<td>0.81</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.70</td>
<td>0.93</td>
<td>0.73</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Figure 1. BM values per file related to the Vdobj

Table 2. Comparison of the three methods concerning Vmod

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-based</td>
<td>0.63</td>
<td>0.73</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td>S-based</td>
<td>0.66</td>
<td>0.83</td>
<td>0.76</td>
<td>0.79</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.64</td>
<td>0.86</td>
<td>0.71</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Figure 2. BM values per file related to the Vmod

Conclusions

• The hybrid model aligns more closely with the correct predictions established by the benchmark set compared to the P-based and the S-based method.
• The hybrid approach outperforms P-based and S-based approach in benchmark match and recall values.

Future work

• Optimise the model as precision, accuracy and F1 score obtain higher values with a P-based approach.
• Enhance the performance of the S-based approach by implementing additional Python rules (negative rules, i.e., rules capable of removing false positive).

References