Funded by the European Union
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

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

Damiano Perri[a], Irene Fioravanti[b], Osvaldo Gervasi[a], Stefania Spina[b]
[a]University of Perugia,
[b]University for Foreigners of Perugia, Italy
damiano.perri, osvaldo.gervasi@unipg.it
irene.fioravanti, stefania.spina@unistrapg.it

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 (Krenn, 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).

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?

The research has been funded by the Italian Ministry of Research (MUR), PRIN: Research Projects of Major National Interest – Call 2022 - Prot. 2022HXZR5E. The title of the project is: DICI-A: A Learner Dictionary of Italian Collocations.

Method

Two types of collocations: **Vdobj** (verb + direct object) and **amod** (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 via 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.

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 **amod** 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]

Results

- Hybrid approach outperforms the P-based and the S-based methods in terms of *recall* and *benchmark match* (BM).
- The P-based method exhibits better *precision* but lower *recall.*
- The S-based method shows lower *precision* but high *recall.*
- All the three methods perform better in detecting amod relations compared to Vdobj.

Table 1. Comparison of the three methods concerning amod

	Accuracy	Recall	Precision	F1	BM
P-based	0.76	0.83	0.90	0.87	83.43%
S-based	0.68	0.88	0.75	0.81	88.25%
Hybrid	0.70	0.93	0.73	0.82	93.37%

Table 2. Comparison of the three methods concerning Vdobj

	Accuracy	Recall	Precision	F1	BM
P-based	0.63	0.73	0.82	0.77	73.33%
S-based	0.66	0.83	0.76	0.79	82.96%
Hybrid	0.64	0.86	0.71	0.78	86.30%

Figure 1. BM values per file related to the amod

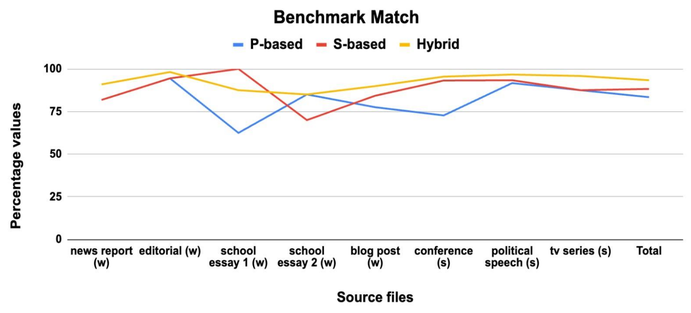
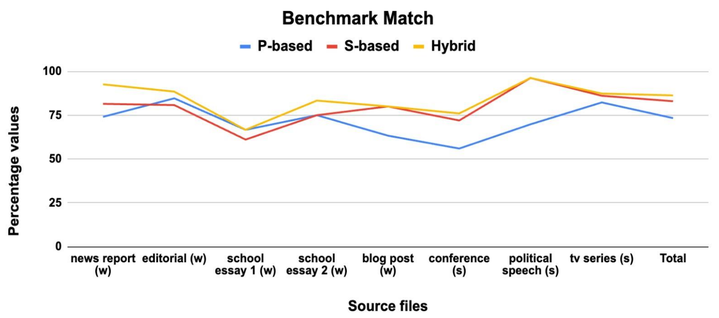


Figure 2. BM values per file related to the Vdobj



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

Castagnoli, S., Lebani, G. E., Lenci, A., Masini, F., Nissim, M., & Passaro, L. C. 2016. Pos-patterns or syntax? Comparing methods for extracting word combinations. In Gloria Corpas Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, Tradulex, Geneve, 116-128.

Evert, S. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis. University of Stuttgart.

Krenn, B. 2000. Collocation mining: Exploiting corpora for collocation identification and representation. In *Proceedings of KONVENS* 2000, Ilmenau, Germany.

Lü, Y., & Zhou, M. 2004. Collocation translation acquisition using monolingual corpora. In *Annual Meeting of the Association for Computational Linguistics*.

Ritz, J. 2006. Collocation extraction: Needs, feeds and results of an extraction system for German. In *Proceedings of the workshop on Multiword-expressions in a multilingual context at the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 41-48.

Seretan, V. 2011. *Syntax-based collocation extraction.* Springer, Dordrecht.

Spina, S. 2014. Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, volume 1, Pisa, Pisa University Press, 354-359.

LREC-COLING 2024