Exploring Non-Verbal Predicates in Semantic Role Labeling: Challenges and Opportunities

Riccardo Orlando [*] **Simone Conia** [*] **Roberto Navigli**
Sapienza NLP Group, Sapienza University of Rome
`{orlando,navigli}@diag.uniroma1.it`
`conia@di.uniroma1.it`

Abstract

Although we have witnessed impressive progress in Semantic Role Labeling (SRL), most of the research in the area is carried out assuming that the majority of predicates are verbs. Conversely, predicates can also be expressed using other parts of speech, e.g., nouns and adjectives. However, non-verbal predicates appear in the benchmarks we commonly use to measure progress in SRL less frequently than in some real-world settings – newspaper headlines, dialogues, and tweets, among others. In this paper, we put forward a new PropBank dataset which boasts wide coverage of multiple predicate types. Thanks to it, we demonstrate empirically that standard benchmarks do not provide an accurate picture of the current situation in SRL and that state-of-the-art systems are still incapable of transferring knowledge across different predicate types. Having observed these issues, we also present a novel, manually-annotated challenge set designed to give equal importance to verbal, nominal, and adjectival predicate-argument structures. We use such dataset to investigate whether we can leverage different linguistic resources to promote knowledge transfer. In conclusion, we claim that SRL is far from "solved", and its integration with other semantic tasks might enable significant improvements in the future, especially for the long tail of non-verbal predicates, thereby facilitating further research on SRL for non-verbal predicates. We release our software and datasets at `https://github.com/sapienzanlp/exploring-srl`.

1 Introduction

Over the years, Semantic Role Labeling (Gildea and Jurafsky, 2002, SRL) – the task of identifying the semantic relations between predicates and their arguments – has attracted continued interest. Enticed by the prospect of acquiring one of the ingredients that might enable Natural Language Understanding (Navigli et al., 2022), the research community has striven to overcome numerous challenges in SRL. As a consequence, not only have automatic systems achieved impressive results on complex benchmarks (Shi and Lin, 2019; Conia et al., 2021), such as CoNLL-2005 (Carreras and Màrquez, 2005), CoNLL-2008 (Surdeanu et al., 2008), CoNLL-2009 (Hajič et al., 2009), and CoNLL-2012 (Pradhan et al., 2012), but SRL has also been successfully leveraged to benefit a wide array of downstream tasks in Natural Language Processing and also Computer Vision, including Machine Translation (Marcheggiani et al., 2018; Raganato et al., 2019; Song et al., 2019), Summarization (Hardy and Vlachos, 2018; Liao et al., 2018), Situation Recognition (Yatskar et al., 2016), and Video Understanding (Sadhu et al., 2021), among others.

Notwithstanding the achievements of previous work, we argue that there is still much to be done before the research community can claim SRL is even close to being "solved". One of the simplest yet erroneous assumptions about SRL is that all predicates – or at least the majority of them – are verbs. Quite the contrary, predicates often manifest themselves as nouns, adjectives, and adverbs. For example, in the sentence "Sensational robbery at the bank during the night: two suspects on the loose!", the word *robbery* is a predicate, as it denotes an action, and its arguments are *sensational* (attribute of the robbery), *at the bank* (location), *during the night* (time), and *two suspects* (agents). We highlight two potential issues in the above example. First, an SRL system that analyzes only verbal predicates cannot identify the nominal event in the sentence and, in turn, its semantic constituents. Second, nominal events like those expressed in the above sentence are far from rare, being commonly found in several settings, such as newspaper headlines, blog titles, short messages, tweets, and

[*] Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2023, pages 12378–12388
July 9-14, 2023 ©2023 Association for Computational Linguistics

dialogues.

Perhaps surprisingly, there is limited work on non-verbal predicates, mostly focused on transferring "knowledge" about verbal predicates to nominal ones (Zhao and Titov, 2020; Klein et al., 2020). The scarcity of studies on non-verbal predicates might be explained by the way in which current datasets for SRL are designed, as they focus primarily on verbal predicates (Daza and Frank, 2020; Tripodi et al., 2021; Jindal et al., 2022). Therefore, any progress on non-verbal predicates is often overshadowed by the predominance of verbal instances, resulting in an incomplete picture of the actual situation. The issue is also exacerbated by the fact that, oftentimes, benchmark results are taken at face value. Instead, carrying out in-depth analyses is fundamental, as neural networks have been found to learn patterns that are different from those of humans, especially in semantic tasks (Maru et al., 2022). In this paper, we perform a reality check and explore non-verbal predicates in English SRL. More specifically, our contributions are as follows:

- We provide an empirical demonstration that state-of-the-art systems are not capable of generalizing from verbal to nominal and adjectival predicate-argument structures (PAS) in PropBank-based SRL;
- We investigate whether other PAS inventories – namely, FrameNet, VerbNet, and VerbAtlas – are better suited for transferring learned patterns across predicate types;
- We introduce a novel, manually-annotated challenge set to evaluate current and future SRL systems on verbal, nominal, and adjectival PAS;
- We analyze possible directions and strategies for prospective work on non-verbal SRL.

2 Challenges

As mentioned above, relying on standard benchmarks does not allow us to properly evaluate the performance of state-of-the-art systems on non-verbal SRL. Cases in point are the CoNLL Shared Tasks: CoNLL-2005 covers only verbal predicates; CoNLL-2009 includes verbal and nominal predicates but makes it difficult to compare them, as they belong to two different inventories, PropBank and NomBank, respectively; CoNLL-2012 and its revision in OntoNotes 5.0 (Pradhan et al., 2022) do not cover adjectival predicates. Therefore, identifying unaddressed challenges, especially in non-verbal SRL, is far from trivial.

	Verbs	**Nouns**	**Adjs**	**Framesets**
CoNLL-2009	1090	1337	0	2427
OntoNotes 5.0	2215	782	3	2490
PB-Examples	5465	1384	1599	7481
PB-Unseen	2457	469	1389	4001

Table 1: Comparison of the coverage of each evaluation benchmark in terms of unique framesets by part of speech (i.e. according to their association with predicate occurrences from the various parts of speech), and total number of part-of-speech independent framesets. PB-Examples and PB-Unseen provide more extensive coverage than CoNLL-2009 and OntoNotes 5.0.

Introducing PB-Examples and PB-Unseen. Since OntoNotes 5.0 – the largest gold evaluation framework for PropBank-based SRL – does not comprehensively evaluate different predicate types, we collect the example sentences provided with each predicate in PropBank 3 (Palmer et al., 2005; Pradhan et al., 2022) to create a new evaluation benchmark, named PB-Examples. This allows us to build a "controlled" benchmark, the first on which we can evaluate the performance of PropBank-based SRL on verbal, nominal, and adjectival PAS.

In Table 1 we report statistics on the coverage of CoNLL-2009, OntoNotes 5.0 and PB-Examples in terms of unique framesets (rightmost column), where the considerably higher frameset coverage of PB-Examples is evident. Compared to its alternatives, PB-Examples covers 7481 unique PropBank framesets against 2490 framesets covered in the OntoNotes test set and 2427 in CoNLL-2009. Moreover, when comparing PB-Examples to OntoNotes, the number of unique framesets used in verbal predicate occurrences is more than double (5465 vs. 2215), whereas it is almost double for nominal occurrences (1384 vs. 782). Adjectival occurrences are essentially missing in OntoNotes (with 3 unique framesets only), while PB-Examples covers 1599. We remark that the same PropBank frameset can be used to annotate predicate occurrences from different parts of speech, which explains why the total number of unique framesets does not correspond to the sum of framesets used for verbal, nominal and adjectival predicate occurrences (second, third and fourth column of Table 1).

Given its considerably higher coverage, PB-

	OntoNotes			PB-Examples				PB-Unseen			
	Verbs	**Nouns**	**V+N**	**Verbs**	**Nouns**	**Adjs**	**V+N+A**	**Verbs**	**Nouns**	**Adjs**	**V+N+A**
Predicates											
CN-22$_{\textbf{verbs}}$	95.4	83.5	94.1	79.1	70.7	54.0	74.7	46.8	34.3	42.8	51.4
CN-22$_{\textbf{nouns}}$	47.6	96.5	53.4	65.6	75.4	59.5	69.7	15.4	29.1	4.2	64.1
CN-22$_{\textbf{verbs + nouns}}$	95.4	96.5	95.6	80.7	80.0	56.4	77.5	51.1	38.5	45.1	53.6
Roles											
CN-22$_{\textbf{verbs}}$	84.7	16.4	80.2	57.8	34.6	25.1	49.6	25.6	6.8	16.5	26.1
CN-22$_{\textbf{nouns}}$	11.2	72.8	16.2	15.1	45.1	5.4	22.1	15.4	29.1	4.2	16.3
CN-22$_{\textbf{verbs + nouns}}$	84.7	76.1	84.1	59.7	59.1	25.6	55.2	28.9	17.8	16.7	28.5

Table 2: F1 scores of CN-22 on the test sets of OntoNotes, PB-Examples, and PB-Unseen, divided by predicate type. These results show that a state-of-the-art system is not capable of "transferring knowledge" from one predicate type to another, e.g., from verbs to nouns or vice versa.

Examples also enables a solid evaluation of an SRL system on over 4000 predicate senses that are not included in OntoNotes 5.0; we call this more challenging testbed PB-Unseen. We report statistics on PB-Unseen in the last row of Table 1.

Cross-type knowledge transfer. Now that we have wide-coverage multi-type SRL datasets, we can test the ability of SRL systems to generalize across types. The main objective of our experiments here is to empirically demonstrate that: i) "knowledge transfer" between predicate types is an unaddressed challenge, and ii) this problem is not apparent in OntoNotes, but becomes evident from PB-Examples and PB-Unseen. To prove these points, we take CN-22 – a state-of-the-art system (Conia and Navigli, 2022) – and study its behavior when trained on the entire OntoNotes (CN-22$_{\textbf{verbs+nouns}}$), only on its verbal structures (CN-22$_{\textbf{verbs}}$), or only on its nominal structures (CN-22$_{\textbf{nouns}}$). The results on the test set of OntoNotes, shown in Table 2, represent the first evidence that even a state-of-the-art SRL system is affected by limited generalization capabilities across predicate types. Indeed, the performance of CN-22$_{\textbf{verbs}}$ drops significantly when evaluated on nominal PAS, from 84.7 to 16.4 points in F1 score on argument labeling, and that of CN-22$_{\textbf{nouns}}$ drops analogously when evaluated on verbal instances, from 72.8 to 11.2 on argument labeling.

One could observe that CN-22$_{\textbf{verbs+nouns}}$, jointly trained on verbal and nominal instances, seems to solve the cross-type transfer problem. However, this is true only because the OntoNotes test set does not feature adjectival structures. Indeed, it is very clear from the results on our PB-Examples and PB-Unseen that the performance of CN-22$_{\textbf{verbs+nouns}}$ does not improve on adjectival PAS compared to CN-22$_{\textbf{verbs}}$ (only +0.5% on PB-Examples and +0.2% on PB-Unseen for argument labeling). Therefore, we can derive that joint learning on two predicate types (i.e. the verbal and nominal ones) does not provide breakthrough improvements on a third predicate type (i.e. the adjectival one). We stress that, in this case, we cannot simply rely on jointly training CN-22 on verbal, nominal, and adjectival instances as, to our knowledge, no training dataset includes adjectival PAS for PropBank-based SRL.

3 Opportunities

In the previous Section, our experiments show that zero-shot knowledge transfer across predicate types is still challenging. We argue that this problem is caused by two main factors. First, PropBank was not designed to aid cross-type knowledge transfer, e.g., the nominal predicate *theft.01* is not linked to its verbal equivalent *steal.01*. Second, recent SRL systems might have limited capability for recognizing common patterns across different predicate types. We conduct an initial investigation of these aspects and discuss some opportunities for improving non-verbal SRL.

The role of the linguistic resource. While PropBank might not be the ideal resource for non-verbal SRL, other inventories – based on different linguistic theories – may provide features that could be helpful to aid knowledge transfer between predicate types. After all, previous studies have already shown that language models leverage different hidden layers depending on the linguistic resource used for SRL (Kuznetsov and Gurevych, 2020; Conia and Navigli, 2022). Here, instead, we take the opportunity to study if there is an inventory whose

	Predicates			Roles		
	P	R	F1	P	R	F1
CN-22 $_{\textbf{PropBank}}$	99.1	96.7	97.9	88.3	88.0	88.1
CN-22 $_{\textbf{FrameNet}}$	99.1	96.7	97.9	89.3	89.5	89.4
CN-22 $_{\textbf{VerbNet}}$	**99.9**	97.4	98.6	**89.8**	89.3	89.5
CN-22 $_{\textbf{VerbAtlas}}$	99.7	**97.7**	**98.7**	89.4	**90.0**	**89.7**

Table 3: Precision (P), Recall (R), and F1 scores of CN-22 on Parallel-SemLink. For each row, we evaluate the performance of the system when trained using the related inventory, e.g., CN-22 $_{\textbf{PropBank}}$ is trained on Parallel-SemLink annotated with PropBank and the results are reported against the test set for the same inventory.

theoretical principles can aid the generalization capability of an existing SRL system on unseen patterns.

We thus evaluate empirically the differences between four different inventories, namely, PropBank, FrameNet (Baker et al., 1998), VerbNet (Schuler and Palmer, 2005), and VerbAtlas (Di Fabio et al., 2019).[1] To do this, we create Parallel-SemLink, a multi-inventory benchmark made up of the subset of OntoNotes from SemLink 2.0 (Stowe et al., 2021), whose predicates and arguments are annotated with PropBank, FrameNet, and VerbNet. We also include VerbAtlas annotations thanks to the inter-resource mapping between VerbNet, WordNet, and VerbAtlas.[2] For each of these inventories, Parallel-SemLink includes a training, a validation, and a test set with 7336, 816, and 906 sentences, respectively.

While we stress that this experimental setting is severely limited since it assumes that all resources can be mapped to each other 1-to-1, it provides a controlled environment for a fair, direct comparison. To study the impact of the inventory, we evaluate our SRL system on each of the linguistic inventories in Parallel-SemLink (CN-22 $_{\textbf{PropBank}}$, CN-22 $_{\textbf{FrameNet}}$, CN-22 $_{\textbf{VerbNet}}$, and CN-22 $_{\textbf{VerbAtlas}}$). The results in Table 3 testify that the linguistic resource of choice plays a role in the results. In particular, we can observe a relative error rate reduction of 38% in predicate sense disambiguation (from 97.9 to 98.7) and 13% in argument labeling (from 88.1 to 89.7) when using VerbAtlas instead of PropBank. This result indicates that higher-level semantic abstractions, such as semantics-based clusters, as available in VerbAtlas thanks to its organization of frames as verbal synset groupings, and cross-predicate role semantics, as adopted in VerbNet and also VerbAtlas, can help a system generalize better on unseen patterns.

	Verbs	Nouns	Adjs	V+N+A
Predicates				
CN-22 $_{\textbf{PropBank}}$	14.5	22.2	27.7	21.7
CN-22 $_{\textbf{VerbAtlas}}$	49.4	17.7	13.5	26.0
Roles				
CN-22 $_{\textbf{PropBank}}$	5.5	2.1	10.8	54.2
CN-22 $_{\textbf{VerbAtlas}}$	47.0	44.2	36.8	42.8

Table 4: F1 scores of CN-22 on Challenge-SRL.

Challenge-SRL. While our multi-inventory SemLink-based dataset provides a preliminary indication of the role of a linguistic inventory, it only includes verbal predicates. To further validate the preliminary results obtained on our multi-inventory SemLink-based dataset, we create a small challenge test set for verbal, nominal, and adjectival SRL, manually annotated with parallel labels for PropBank, the most popular inventory, and VerbAtlas, the most promising inventory (cf. Table 3). This new test set is particularly challenging, as it features only PAS that do not appear in OntoNotes. Therefore, Challenge-SRL makes it possible to measure the capability of an SRL system to generalize i) across predicate types, and ii) on the long tail of predicate senses.

To construct Challenge-SRL, we randomly selected a total of 288 sentences – 96 sentences for each predicate type – from PB-Unseen. We then asked three expert annotators to independently annotate each sentence with predicate senses and their semantic roles. The annotation process was carried out in two phases: first, each person annotated each sentence independently, resulting in a disagreement of 32%; then, the annotators discussed and resolved their disagreements, if possible, reducing them to 6%. Overall, Challenge-SRL includes 1898 predicate-argument pairs.

As we can see from Table 4, Challenge-SRL confirms our preliminary experiments, macroscopically magnifying the differences between PropBank and VerbAtlas. First, we observe that VerbAtlas is significantly better in predicate sense disambiguation for verbal instances (49.5 vs. 14.5 in F1 score) but worse for nominal and adjectival ones

[1] Appendix A provides an overview of the inventories.

[2] Appendix C provides further details on our mapping procedure.

	Verbs	Nouns	Adjs	V+N+A
CN-22 $_{\textbf{SemLink}}$	6.2	6.2	3.1	5.2
CN-22 $_{\textbf{OntoNotes}}$	49.4	5.2	10.2	26.0
WSD $_{\textbf{baseline}}$	46.7	32.7	3.8	31.7
Oracle $_{\textbf{SL+WSD}}$	58.9	37.2	9.3	31.4
Oracle $_{\textbf{ON+WSD}}$	**60.5**	**41.6**	**25.6**	**41.5**

Table 5: F1 scores on predicate disambiguation of CN-22 and a WSD system on Challenge-SRL. The scores of Oracle $_{\textbf{SL+WSD}}$ (Oracle $_{\textbf{ON+WSD}}$) are obtained by picking the best prediction between WSD $_{\textbf{baseline}}$ and CN-22 $_{\textbf{SemLink}}$ (CN-22 $_{\textbf{OntoNotes}}$).

(22.2 vs. 17.7 and 27.7 vs. 13.5, respectively). This is mainly because VerbAtlas was not designed for non-verbal SRL and, therefore, it does not provide a lemma-to-sense dictionary to restrict the possible frames of nominal and adjectival predicates. Second, VerbAtlas significantly outperforms PropBank on argument labeling of verbs (47.0 vs. 5.5 in F1 score), nouns (44.2 vs. 2.1), and adjectives (36.8 vs. 10.8). We argue that this is largely due to the adoption in VerbAtlas of cross-frame semantic roles that are coherent across frames, which allows the system to leverage other predicates seen at training time with similar structures.

Leveraging Word Sense Disambiguation. Finally, we carry out a preliminary exploration of possible directions that could aid non-verbal SRL in the future. While SRL research has not dealt with non-verbal semantics, other areas have investigated semantics for different parts of speech, and one of these is Word Sense Disambiguation (WSD). More specifically, WSD is the task of assigning the most appropriate sense to a word in context according to a predefined sense inventory (Bevilacqua et al., 2021). It is easy to notice how this task resembles predicate sense disambiguation in SRL, the only difference being that WSD is not limited to predicates, as it aims to disambiguate every content word. Therefore, we believe that WSD is an interesting candidate to explore whether a different disambiguation task can help to improve the generalization capability of an existing SRL system on Challenge-SRL, i.e., on predicate-argument structures that the SRL system did not see at training time.

To investigate the effect of WSD on SRL, we start by leveraging the fact that VerbAtlas frames are clusters of WordNet synsets. Therefore, we map each synset predicted by AMuSE-WSD (Orlando et al., 2021, 2022),[3] a state-of-the-art off-the-shelf WSD system, to a VerbAtlas frame, and compare them to the prediction of our SRL system. Table 5 shows the performance of AMuSE-WSD on predicate sense disambiguation (WSD$_{\text{baseline}}$). Interestingly, we observe that a simple WSD baseline can strongly outperform an SRL system when training data is scarce. Indeed, AMuSE-WSD surpasses CN-22 $_{\textbf{SemLink}}$ in each predicate type (46.7 vs 6.2, 32.7 vs 6.2, 3.8 vs 3.1, for verbs, nouns and adjectives, respectively), and CN-22 $_{\textbf{OntoNotes}}$ in nominal predicates, with an overall improvement of +5.7 (31.7 vs 26.0) over the best performing SRL system.

Most interestingly, if we employ an oracle to pick the best prediction between the WSD baseline and our best SRL system, we notice a further improvement (41.5% vs. 26.0%), demonstrating that current state-of-the-art SRL systems can still benefit from explicit lexical semantics. We hypothesize that tighter integration of the two tasks may lead to even better improvements in generalization capabilities.

4 Conclusion and Future Work

In this paper, we carried out a reality check and demonstrated that, despite impressive results on standard benchmarks by state-of-the-art systems, SRL is still far from “solved”. Indeed, thanks to a carefully-designed set of experiments and the introduction of novel, manually-curated, wide-coverage benchmarks, we showed that current SRL systems possess inadequate capabilities for transferring knowledge between predicate types.

Our analyses pointed out that we can address this limitation by working in two directions: leveraging the intrinsic characteristic of frameset resources, including semantics-based clusters and cross-predicate role semantics, and tighter integration of other semantics-based tasks, such as Word Sense Disambiguation, into SRL.

We hope our work will be a stepping stone for innovative research on high-performance SRL systems for non-verbal predicate-argument structures, a problem that still needs extensive investigation. For this reason, we release our software and datasets at `https://github.com/sapienzanlp/exploring-srl`.

[3] `https://nlp.uniroma1.it/amuse-wsd/`

Limitations

Part of our analyses and experiments is based on our Parallel-SemLink dataset, which provides parallel annotations for PropBank, FrameNet, VerbNet, and VerbAtlas. We take the opportunity to remark that this is a constrained setting, as these resources cannot be mapped 1-to-1 without losing information. As such, this setting may not provide the full picture of how these resources compare against each other. However, we also believe that a setting like this can at least provide an intuitive idea of the role of a linguistic resource in cross-inventory generalization. Creating novel benchmarks that can better compare the role of different linguistic resources is certainly a direction for future work that may provide novel insights into verbal and non-verbal SRL.

Another limitation of our work is the small size of Challenge-SRL. Even though Challenge-SRL contains only about 300 sentences, it features almost 2000 predicate-argument pairs, and this is a number that is sufficient to show the inability of a current state-of-the-art system to generalize across predicate types. We acknowledge that a larger benchmark may have provided further insights. However, we also note that, in our case, increasing the number of annotations would hardly have brought us to a different conclusion, especially given the large differences in performance among the model configurations that we evaluated.

Finally, we stress that our experiments on integrating a simple WSD baseline into an SRL system do not provide a definitive answer on whether more complex integrations may lead to improved results. Instead, our intention is to support the claim that SRL is still far from being "solved", as knowledge from other tasks can still hypothetically bring benefits to an existing SRL system, especially when the size of the training data is small.

Ethics Statement

We release all the new datasets we produce under an open license. However, some of the datasets mentioned and used in our paper are not openly available, e.g., CoNLL-2009 and OntoNotes 5.0. We acknowledge the fact that such datasets may become unavailable at a later moment, as their distribution is not under our control.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

The last author gratefully acknowledges the support of the PNRR MUR project PE0000013-FAIR.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, page 86–90, USA. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan. Association for Computational Linguistics.

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–351, Online. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632, Dublin, Ireland. Association for Computational Linguistics.

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling.

In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

Hardy Hardy and Andreas Vlachos. 2018. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773, Brussels, Belgium. Association for Computational Linguistics.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal Proposition Bank 2.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France. European Language Resources Association.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182, Online. Association for Computational Linguistics.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Diego Marcheggiani, Jasmijn Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of Word Sense Disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Roberto Navigli, Edoardo Barba, Simone Conia, and Rexhina Blloshmi. 2022. A tour of explicit multilingual semantics: Word sense disambiguation, semantic role labeling and semantic parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 35–43, Taipei. Association for Computational Linguistics.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Riccardo Orlando, Simone Conia, Stefano Faralli, and Roberto Navigli. 2022. Universal semantic annotator: the first unified API for WSD, SRL and semantic parsing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2634–2641, Marseille, France. European Language Resources Association.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-bettner, and Martha Palmer. 2022. PropBank comes of Age—Larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual Semantic Role Labeling for Video Understanding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5585–5596. ISSN: 2575-7075.

Karin Kipper Schuler and Martha Palmer. 2005. Verbnet: a broad-coverage, comprehensive verb lexicon.

Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Kevin Stowe, Jenette Preciado, Kathryn Conger, Susan Windisch Brown, Ghazaleh Kazeminejad, James Gung, and Martha Palmer. 2021. SemLink 2.0: Chasing Lexical Resources. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 222–227, Groningen, The Netherlands (online). Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177, Manchester, England. Coling 2008 Organizing Committee.

Rocco Tripodi, Simone Conia, and Roberto Navigli. 2021. UniteD-SRL: A unified dataset for span- and dependency-based multilingual and cross-lingual Semantic Role Labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2293–2305, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mark Yatskar, Luke S. Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5534–5542. IEEE Computer Society.

Yanpeng Zhao and Ivan Titov. 2020. Unsupervised Transfer of Semantic Role Models from Verbal to Nominal Domain.

A Inventories

In this paper, we evaluate empirically how SRL systems are influenced by the different linguistic inventories employed. We tested four popular inventories, namely PropBank, FrameNet, VerbNet, and VerbAtlas. Each of these inventories features different characteristics, which we summarize briefly here.

PropBank PropBank (Palmer et al., 2005) enumerates the senses of each predicate lemma, e.g., *eat.01*, *eat.02*, etc., and defines semantic roles (ARG0-ARG5) that are specific to each predicate sense, e.g., the meaning of ARG2 in *eat.01* differs from that of *eat.02*.

FrameNet FrameNet (Baker et al., 1998) groups predicates that evoke similar actions in semantic frames, e.g., the frame *Ingestion* includes eating, feeding, devouring, among others; each frame can have frame-specific roles, e.g., INGESTOR and INGESTIBLE.

VerbNet VerbNet (Schuler and Palmer, 2005) defines classes of verbs with similar syntactic patterns, e.g., eating and drinking belong to *Eat-39.1-1*; all verb classes share a set of thematic roles, e.g., AGENT and PATIENT.

VerbAtlas VerbAtlas (Di Fabio et al., 2019) clusters WordNet (Miller, 1992) synsets into coarse-grained frames, similar to FrameNet, and adopts a common set of thematic roles for all frames, similar to VerbNet.

B Parallel-SemLink

In this Section, we provide further details on the construction process of Parallel-SemLink. We leverage the data distributed as part of SemLink 2.0 (Stowe et al., 2021), which includes instances from OntoNotes 5.0 annotated with PropBank, FrameNet, and VerbNet. We select the subset of the instances that have a corresponding annotation in all three inventories. In addition, we also include VerbAtlas annotations through the inter-resource mapping between VerbNet, WordNet, and VerbAtlas. To convert the predicate senses, we employ the mapping from VerbNet to WordNet included in the Unified Verb Index (UVI)[4] project: since a VerbAtlas frame is a cluster of WordNet synsets, we associate a VerbNet class with a VerbAtlas frame

[4] `https://uvi.colorado.edu/`

through their corresponding synset. Additionally, we also extend the VerbAtlas annotations to include argument roles. Given that both VerbNet and VerbAtlas adopt a similar set of thematic roles, we manually map all the VerbNet roles to their corresponding VerbAtlas ones and convert the argument annotations accordingly.

C Mapping Nouns to VerbAtlas Frames

Since VerbAtlas was originally designed only as a verbal inventory, its frames contain only verbal WordNet synsets. To expand its coverage and include nominal predicates, we propose a method for deriving nominal predicates from the verbal ones already included. The method leverages WordNet (Miller, 1992), a lexical database that contains a wealth of information about word senses and their relationships. Specifically, we use the "hypernym" and "derivationally related forms" relations in WordNet to identify nominal word senses that are semantically related to a verbal predicate in VerbAtlas. Informally, to be included in our expanded version of VerbAtlas, a nominal word sense must meet the following criteria:

1. It must have a "hypernym" that belongs to the top-100 most frequent nominal senses related to *event.n.01*, i.e., event as in "something that happens at a given place and time".

2. It must be semantically related – "derivationally related forms" related – to a verbal predicate included in a VerbAtlas frame.

This approach allows us to identify a large number of nominal word senses that are semantically related to a verbal predicate in VerbAtlas. Therefore, we assign these nominal word senses to the same VerbAtlas frame as their related verbal predicates. In total, we are able to cluster 5334 nominal word senses, significantly expanding the coverage of VerbAtlas to include both verbal and nominal predicates. We release this mapping together with the rest of our software and datasets.

D Mapping Adjectives to VerbAtlas Frames

We follow a similar strategy to also include adjectival predicates in VerbAtlas. This time, we rely on the "pertainyms", "similar to", and "derivationally related forms" relations to connect adjectival word senses in WordNet to VerbAtlas frames. More specifically, we include each adjectival word sense that satisfies at least one of the following conditions:

- It must be "derivationally related" or "pertaining" to a noun or verb sense that is already included in VerbAtlas;

- It must be "similar to" another word sense that is in turn "derivationally related" to a predicate in VerbAtlas.

We then assign these adjectival word senses to the same VerbAtlas frame as their related verbal and nominal predicates. As a result, we are able to include 2968 adjectival predicates in VerbAtlas. We release this mapping together with the rest of our software and datasets.

E License

We release our data under the Creative Commons Attribution Share-Alike (CC-BY-SA) license.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2,3

☑ B1. Did you cite the creators of artifacts you used?
1,2,3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
11

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Among the existing artifacts that we employ in our work, we use datasets and models that have been originally designed for Semantic Role Labeling and we continued to use them according to their original intended usage. All the datasets and models that we create are based on Semantic Role Labeling resources and we use them for Semantic Role Labeling

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Our data was randomly selected from existing sources to ensure the same distribution.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

2,3,4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We didn't report any of that information because it was not relevant to our work.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We used already existing systems to carry out our experiments.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
2,3,4

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Our annotations were based on the guidelines of PropBank and VerbAtlas, which are cited in the paper.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We will add this information in the camera-ready in case of acceptance. We did not include this information at submission time to not invalidate the anonymity of the paper.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our university does not have a board for this kind of work.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The annotators did not agree to disclose this information.