Common Sense Bias in Semantic Role Labeling

Heather Lent and **Anders Søgaard**
University of Copenhagen, Denmark
{hcl, soegaard}@di.ku.dk

Abstract

Large-scale language models such as ELMo and BERT have pushed the horizon of what is possible in semantic role labeling (SRL), solving the out-of-vocabulary problem and enabling end-to-end systems, but they have also introduced significant biases. We evaluate three SRL parsers on very simple transitive sentences with verbs usually associated with animate subjects and objects, such as *Mary babysat Tom*: a state-of-the-art parser based on BERT, an older parser based on GloVe, and an even older parser from before the days of word embeddings. When arguments are word forms predominantly used as person names, aligning with common sense expectations of animacy, the BERT-based parser is unsurprisingly superior; yet, with abstract or random nouns, the opposite picture emerges. We refer to this as *common sense bias* and present a challenge dataset for evaluating the extent to which parsers are sensitive to such a bias. Our code and challenge dataset are available here: github.com/coastalcph/comte

1 Introduction

Semantic role labeling (SRL) refers to a shallow semantic dependency parsing that returns predicate-argument structures for input sentences; see Figure 1. Modern-day SRL systems, like most other NLP technologies, rely heavily on large-scale language models. Such language models are extremely useful for generalizing to out-of-vocabulary items, making subtle syntactic distinctions, and for capturing a range of lexical ambiguities; but they also introduce notable biases.

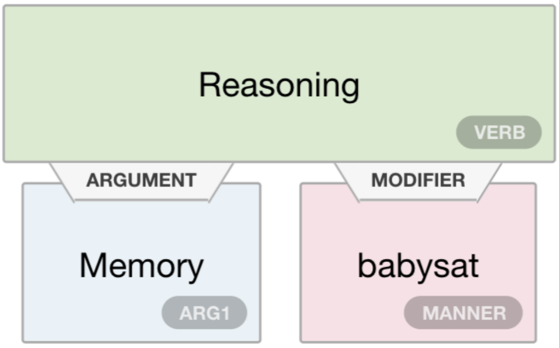

Figure 1: The (incorrect) analysis of *Memory babysat Reasoning* by Shi and Lin (2019).

Previous work has shown that SRL systems exhibit demographic biases (Zhao et al., 2017); we focus on a form of belief bias (Sternberg and Leighton, 2004), which we will refer to as *common sense bias*, reflecting how language models encode conventional associations, which in many ways are indistinguishable from common sense (Trinh and Le, 2019). While demographic biases can lead to discrimination against under-represented demographics, belief biases can lead to discrimination against rare events; or, more precisely, lead SRL systems to err on sentences that express unlikely states of affairs. This is what belief biases refer to in cognitive science (Sternberg and Leighton, 2004): human preferences for conclusions that align with values, beliefs, and prior knowledge. Belief biases in models can, like demographic biases, exacerbate societal challenges, e.g., anomaly detection, and also correlate with demographics, since groups differ in how much they engage with counterfactual and fictitious contents.

We compare the errors of a modern, competitive SRL system (Shi and Lin, 2019), based on BERT (Devlin et al., 2019), and show how it, unlike earlier SRL systems, suffers from common sense bias: When confronted with sentences that, when read literally, express unlikely states of affairs, it can ignore obvious cues and produce false predicate-argument structures even for very simple sentences. The sentence in Figure 1, for example, can be understood as expressing that the abstract concept of *Memory* babysat the abstract concept of *Reasoning*. The literal reading represents an unlikely state

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 114–119
November 11, 2021. ©2021 Association for Computational Linguistics

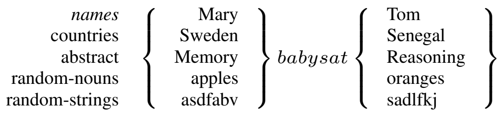

Figure 2: Examples of transitive sentences with person names, country names, abstract nouns, (randomly chosen) plural common nouns, or random strings as arguments. Person names, and to some degree country names (which are often personified (Wang, 2020)), align with expectations of animacy.

Ref	Model	F_1
Björkelund et al. (2009)	MST/MIRA	0.803
Stanovsky et al. (2018)	LSTM/GloVe	0.823
Shi and Lin (2019)	BERT	0.888

Table 1: The three SRL systems used below and their performance on the CoNLL 2005 benchmark

of affairs, since abstract concepts generally do not have the capacity of babysitting. Obviously, this does not prevent language users from uttering the sentence, and it is, for most of us, not hard to make sense of it: The sentence, for example, could mean something like *memory assists reasoning*. Many similar sentences can be found in the wild, e.g., *the US babysits Israel* (from `cnn.com`) or *Love bodyslams you* (from `quizlet.com`). Other sentences express unlikely states of affairs, not because of linguistic creativity, but because they refer to possible worlds, not ours, for scientific, literary, political or other reasons. We believe it is critical that SRL parsers should be robust to such variation, but our experiments show that while SRL performance numbers have gone up dramatically in recent years, parsers seem to have become more sensitive to it.

Contributions We present an error analysis of three very different SRL parsers for English: the supervised, log-linear, quadratic-time parser proposed in Björkelund et al. (2009); the supervised, deep, linear-time parser proposed in Stanovsky et al. (2018), based on GloVe embeddings (Pennington et al., 2014) and recurrent networks (Hochreiter and Schmidhuber, 1997); and the self-supervised (and supervised), deep, linear-time parser proposed in Shi and Lin (2019), based on BERT (Devlin et al., 2019). Instead of evaluating these models on standard benchmarks of newspapers, where predicate-argument structures already align with the 'beliefs' of BERT, we evaluate the systems on randomly generated transitive sentences of the form NP-V-NP, with V expressed by verbs strongly associated with A_0-V-A_1 frames, and the NPs expressed by proper nouns, abstract nouns or plural common nouns. From these experiments, we show that (a) the SRL systems considered here frequently err on such sentences; (b) the SRL error distribution across verb lemmata is uncorrelated with the errors of a dependency parser; (c) what pairs of NP semantic categories lead to errors for what verbs; and (d) how the BERT-based system suffers from common sense bias. Finally, we create a 1000-sentence challenge dataset for probing SRL for common sense bias. Our error analyses paint a complementary, yet entirely different picture of what SRL systems can and cannot, compared to previous work (He et al., 2017; Strubell et al., 2018), which has focused on long-distance dependencies and the need for syntax.

2 Semantic Role Labeling Systems

Björkelund et al. (2009) combine three logistic regression classifiers with beam search and a global reranker: the first classifier identifies predicates, the second their arguments, and the third labels the semantic dependencies between predicates and their arguments. The system relies on a POS tagger and a syntactic dependency parser to generate features for the classifiers. This system had the second-best performance in the CoNLL 2009 Shared Task. Stanovsky et al. (2018) rely on a standard recurrent architecture. They use GloVe embeddings (Pennington et al., 2014), in conjunction with embeddings from a POS tagger, and stack bidirectional LSTM layers (Hochreiter and Schmidhuber, 1997) on top of the embedding layer. The representation at each time-step is passed to a classifier, which directly predicts the output label for that time-step. Unlike Björkelund et al. (2009), they do not rely on search over possible output combinations. Shi and Lin (2019) also do not rely on search, but reduce SRL to two-stage sequence labeling, both stages pretrained with BERT-large (Devlin et al., 2019); first identifying predicates, then arguments, while conditioning on the predicates.

3 Coarse-Grained Error Analysis

In our error analysis, we focus on simple three-word sentences that consist of a noun, a transitive verb, and a noun. The transitive verbs are hand-picked to exhibit strong preferences for animate subjects and objects, low ambiguity, and predom-

Verb	Error	A_0VA_2	A_1VA_2	…V	Expl
fails	0.898	0.006	**0.758**	0.000	Syntax
calls	0.528	**0.467**	0.020	0.020	
trips	0.356	0.007	0.000	**0.128**	POS
tips	0.875	0.010	0.010	**0.687**	
bodyslams	0.373	0.065	0.052	0.034	?
babysits	0.212	0.048	0.072	0.035	

Table 2: Error rates and most frequent error types for common verbs in their present and past tense forms, in simple SOV constructions, e.g., *John calls Mary*. All numbers are for Shi and Lin (2019). Bold-faced error types most frequent (of the four presented here). The verbs *bodyslams* and *babysits* are used in our experiments, because (a) they have strong selectional restrictions for animate subjects and objects, (b) they predominantly realize A_0 and A_1 as subjects and objects (unlike *fails* and *calls*), and (c) while all English verbs tend to have noun readings, the verb readings are far more frequent (unlike for *trips* and *tips*).

Error	Björkelund et al. (2009)	Stanovsky et al. (2018)	Shi and Lin (2019)
names	0.158	**0.341**	0.077
countries	0.183	**0.505**	0.030
abstract	0.174	**0.353**	0.133
random-nouns	0.188	0.287	**0.310**
random-strings	**0.997**	0.172	0.313

Table 3: **Main results:** Error rates of three SRL systems across transitive sentences with person names in subject and object positions, versus country names, abstract nouns, (randomly chosen) plural common nouns, or random strings in those positions

Tajikistan bodyslams Maldives	Lebanon bodyslammed Netherlands
Myanmar bodyslammed Andorra	Bangladesh bodypaints Peru
Luxembourg bodypainted Andorra	Kazakhstan bodypainted Guinea
Bangladesh combed Turkey	Bangladesh manicured Swaziland

Table 4: Simple sentences on which Stanovsky et al. (2018) and Shi and Lin (2019) both err. Björkelund et al. (2009), in contrast, assigns correct parses to all of these. Try yourself: `barbar.cs.lth.se:8081/`

inantly realize their agents (A_0) as subjects, and their second argument (A_1) as objects. The error analysis consists of comparing performance across different types of subjects and objects and comprises examples such as those in Figure 2. The arguments exhibit various degrees of animacy associations, aligning more or less with common sense expectations. We obtain the names from the NAMES library,[1] the country names from WorldMap,[2] the abstract nouns from YourDictionary,[3] and common nouns from the Princeton WordNet.[4]

We assume a correct semantic parse associates subject with A_0 and object with A_1 (of the predicate introduced by the verb). This is obviously not true for all verbs (Palmer et al., 2005; Hovy et al., 2006). In Table 2, we list verbs that frequently associate subjects and objects with other arguments (*fails* and *calls*), as well as verbs that are very ambiguous and easily mistaken for nouns (*trips* and *tips*). Both phenomena are reflected in the distribution of analyses for Shi and Lin (2019). While much can be said about these verbs, our main contribution here is highlighting the role of common sense bias in some SRL parsers, and we thus focus on verbs where we can safely assume a A_0VA_1 reading is correct (such as *babyslams* and *babysits*).[5]

Error analysis results are presented in Table 3. If performance drops considerably below the performance label with names or countries, when using abstract nouns, randomly sampled nouns, or simply random strings, as arguments, this suggests a common sense bias, seen very strongly with Shi and Lin (2019). Björkelund et al. (2009), in contrast, exhibits near-uniform performance across the different sets of arguments. Since the parser has no strategy to deal with out-of-vocabulary items, it exhibits worse performance on random strings.[6] Stanovsky et al. (2018), surprisingly, seems extremely sensitive to country name arguments,[7] and performance oddly improves with random strings arguments. Since these are out-of-vocabulary, the parser probably drops back to a default strategy. Notably, Shi and Lin (2019) does well on country names, there are plenty of examples that Stanovsky et al. (2018) and Shi and Lin (2019) get wrong, but that Björkelund et al. (2009) get right; see Table 4 for examples.

[1] `https://pypi.org/project/names/`

[2] `http://worldmap.harvard.edu/`

[3] `https://examples.yourdictionary.com/examples-of-abstract-nouns.html`

[4] `https://wordnet.princeton.edu/`

[5] The six verb lemmata we use are: *bodyslam*, *bodypaint*, *comb*, *manicure*, *elbow*, and *babysit*.

[6] Björkelund et al. (2009) near-consistently analyze these as intransitive with the first two words making up A_1.

[7] We found no explanation for Stanovsky et al. (2018)'s poor performance with country name arguments.

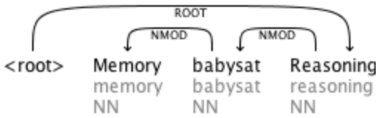

Figure 3: Parse tree in Björkelund et al. (2009) for *Memory babysat Reasoning*.

Comparison with Dependency Parser Errors While only one of the parsers (Björkelund et al., 2009) relies on input features from a syntactic parser, it is tempting to think that, in line with previous error analyses of SRL systems (He et al., 2017; Strubell et al., 2018), the error distribution can be explained by syntactic ambiguities and resulting syntactic errors. This, perhaps unsurprisingly, turns out to reliably explain the error distribution observed with Björkelund et al. (2009). See Figure 3 for the syntactic parse on *Memory babysits Reasoning*, on which the log-linear parser fails to deliver any SRL analysis, interpreting the three-word sentence as a nominal compound. For the neural parsers, there is no correlation, however. We ran a syntactic parser (Dozat et al., 2017) on our three-word sentences and correlated errors across lemmata. We observed a small, but *negative* correlation between error rates.

4 Fine-Grained Error Analysis

Multitude of Errors Our first observation is that across all verb lemmata, the parser in Shi and Lin (2019) produces *many* different output trees, depending on the argument word forms. For some lemmata, the error distribution is near-uniform across 15-20 outputs. It is well-established in SRL that infrequent contexts lead to low confidence (Chen et al., 2011), explaining why common sense bias leads to a multitude of errors.

Morphosyntactic Ambiguity While parsing errors do not correlate with errors of Shi and Lin (2019) (§3), the SRL system seems to be sensitive to part-of-speech ambiguity. It errs, for example, on *Insomnia trips jaywalking*, but not on *Insomnia tripped jaywalking*, presumably because *trips* is (on its own) ambiguous.[8] Sensitivity to such ambiguities disappears when aligning with common sense: The parser does not err on *Mary trips John* or *Mary likes jaywalking*. The same ambiguity leads to error in *London trips John.*, but not in *London tripped John.* With the even more frequent surname of *Washington*, the effect disappears, and Shi and Lin (2019) get both verb forms right.

5 COMTE: A Test of Common Sense Bias

Our challenge dataset[9] COMTE consists of 1,000 simple, three-word sentences with the same gold analysis: the second word is the predicate, the first word its A_0, the last word its A_1. The predicates are sampled at random from a list of six carefully selected verbs (see §3) that select for animate subjects and objects and consistently prefer these to be A_0 and A_1. As before, we combine the verbs with names, countries, abstract nouns, plural common nouns, and random strings. The sentences were simply the first 1,000 sentences that we sampled this way, with 200 sentences in each category (names, countries, etc.) – and which satisfied a simple criterion: Neither Shi and Lin (2019) nor (Stanovsky et al., 2018) would get it right. COMTE, in other words, consists of 1,000 trivial sentences that two competitive SRL parsers failed to parse correctly.

What can COMTE be used for? Obviously, it can not be used to fine-tune parsers on, for example. It would take only a few examples to learn what is going on in the data, and training would likely lead to over-fitting. COMTE can also not be used to derive parsing performance figures that tell us much about the performance of parsers in the wild. The 1,000 sentences should, in our view, be thought of as a single probe into the degree to which a parser is sensitive to common sense bias. A parser should rarely err on the examples in the challenge dataset: They are all trivially simple, and while some argument words can be ambiguous, the verbs so strongly select for simple A_0VA_1 frames that parsers should unambiguously prefer this reading. If they don't, this is a sign they struggle with simple transitive sentences, like Stanovsky et al. (2018), or that they are prone to common sense bias, like Shi and Lin (2019). In order to quantify the degree to which the effect can be attributed to common sense bias, performance with *names* can be used as a baseline: If performance is much better for names than for some of the other categories, like with Shi and Lin (2019), this is an indicator of common sense bias.

[8] This is orthogonal to the ambiguity of *jaywalking*; see Padó et al. (2008) for the analysis of nominal predicates.

[9] Our dataset differs from previous challenge datasets for mixed language (Pal and Sharma, 2019), chat (Rachman et al., 2018), etc., in being synthetic.

6 Acknowledgments

We thank the anonymous reviewers for their useful feedback. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant Agreement No 801199, as well as from a Google Focused Research Award.

References

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado. Association for Computational Linguistics.

Chenhua Chen, Alexis Palmer, and Caroline Sporleder. 2011. Enhancing active learning for semantic role labeling via compressed dependency trees. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 183–191, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Sebastian Padó, Marco Pennacchiotti, and Caroline Sporleder. 2008. Semantic role assignment for event nominalisations by leveraging verbal data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 665–672, Manchester, UK. Coling 2008 Organizing Committee.

Riya Pal and Dipti Sharma. 2019. A dataset for semantic role labelling of Hindi-English code-mixed tweets. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 178–188, Florence, Italy. Association for Computational Linguistics.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Valdi Rachman, Rahmad Mahendra, Alfan Farizki Wicaksono, Ahmad Rizqi Meydiarso, and Fariz Ikhwantri. 2018. Semantic role labeling in conversational chat using deep bi-directional long short-term memory networks with attention mechanism. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Robert Sternberg and Jacqueline Leighton. 2004. *The Nature of Reasoning*. Cambridge University Press.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Trieu H. Trinh and Quoc V. Le. 2019. Do language models have common sense?

Yongqi Wang. 2020. The metaphoric and metonymic use of country names in economic newsa corpus-based analysis. *Chinese Journal of Applied Linguistics*, 43(4):439 – 454.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.