Bridging Perception, Memory, and Inference through Semantic Relations[*]

Johanna Björklund[†] **Adam Dahlgren Lindström**[‡] **Frank Drewes**
Department of Computing Science
Umeå University, Sweden
{johanna,dali,drewes}@cs.umu.se

Abstract

There is a growing consensus that surface form alone does not enable models to learn meaning and gain language understanding. This warrants an interest in hybrid systems that combine the strengths of neural and symbolic methods. We favour triadic systems consisting of neural networks, knowledge bases, and inference engines. The network provides *perception*, that is, the interface between the system and its environment. The knowledge base provides explicit *memory* and thus immediate access to established facts. Finally, *inference* capabilities are provided by the inference engine which reflects on the perception, supported by memory, to reason and discover new facts. In this work, we probe six popular language models for semantic relations and outline a future line of research to study how the constituent subsystems can be jointly realised and integrated.

1 Introduction

Recent works (Bender and Koller, 2020; Bender et al., 2021) postulate that it is impossible to learn meaning from surface form alone, and express concerns about what is perceived as an over-reliance on large-scale pretrained neural networks. This line of thought supports the interest in hybrid systems that amalgamate elements from complementary learning paradigms (see, e.g., (Pearl, 2019; Wang et al., 2019; Hohenecker and Lukasiewicz, 2020; van Bekkum et al., 2021)). In (Dahlgren et al., 2021), we argue that this calls for an explicit distinction to be made between the faculties of perception, memory, and inference. We therefore promote the development of systems that consist of subsystems with responsibilities corresponding to the three faculties. Such future systems would thus consist of a perception component realised by a neural network, a component that provides explicit memory in the form of a knowledge base, and a third one performing symbolic inference, that is, rule-based reasoning.

We suggest to study how the subsystems can be aligned so for a seamless information flow between them. We view it as particularly important that (i) the network and the knowledge base together yield a consistent treatment of semantic relations and (ii) training takes the knowledge base into account, so that the resulting embeddings are consistent with established facts. Our conceptual discussion is complemented by a preliminary empirical evaluation of six popular English language models, which we subject to linear probes to test their abilities to capture central semantic relations.

After a brief discussion of related work in Section 2, Section 3 discusses the role of semantic relations in the context of our envisioned triad system while Section 4 and 5 of this paper complement our conceptual discussion with a preliminary empirical evaluation of the chances to achieve (i) by probing six popular language models with respect to a semantic relation learning task.

2 Related work

There is a rapidly growing literature on relation extraction and hybrid systems. Petroni et al. (2019) observe that language models such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) are imprinted with large amounts of common sense and factual knowledge during training. If this information can be reliably extracted then, they argue, word embeddings could find a new use as knowledge bases. To test the practicality of this approach, they consider a knowledge extraction task where a language model is given a sentence containing a subject word x and a relation $\mathcal{R}$, but where the object word y has been removed, and the model should guess the missing y (i.e., rank the vocabu-

[*] The authors are given in alphabetical order.

[†] The work of this author is partially funded by the Swedish Research Council, Grant number 2020-03852, and the Wallenberg Autonomous Systems and AI program.

[‡] Contact Author

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9136–9142
November 7–11, 2021. ©2021 Association for Computational Linguistics

lary words) based on the fact that x and y are in the relation $\mathcal{R}$. The sentences are generated based on manually constructed templates, one per relation. For example, to the relation *birth-place*, they use the template "$\langle$subject$\rangle$ was born in $\langle$blank$\rangle$" and instantiate it to "Dante was born in $\langle$blank$\rangle$". The most important baselines are two variations of the relation extraction model by Sorokin and Gurevych (2017). Key findings are that language models appear to be better at learning one-to-one relations, whereas the relation extraction models are better at picking out many-to-many relations. Petroni et al. also find that the choice of template has an impact on the performance of the language models, and point this out as an item for future work.

Bouraoui et al. (2020) pick up this thread and propose a method for extracting good template sentences from BERT, and using these to fine-tune BERT so as to improve its performance on relation extraction. For a target binary relation $\mathcal{R}$ (represented as a set of ordered pairs) and a sample of pairs $R \subseteq \mathcal{R}$, they filter the training data for sentences expressing that x and y, with $(x, y) \in R$, have the relation $\mathcal{R}$, and which would still be natural if x and y where simultaneously replaced by some other $(x', y') \in R$. Finally, they fine-tune a language model to predict, from an instantiation of one of the remaining sentences with a pair (x'', y''), whether $(x'', y'') \in \mathcal{R}$. The most relevant aspect of this work for the present effort is the evaluation of the Bigger Analogy Test Set (also known as BATS) which contains 40 relations with 50 instances per relation (Gladkova et al., 2016). Bouraoui et al. (2020) report a mixed performance on the type of semantic relations considered here, namely hypernyms and hyponyms.

Additional methods for choosing template sentences are proposed by Jiang et al. (2020) who, similar to Bouraoui et al. (2020), mine the training data for suitable sentences. A dependency analysis on candidate sentences makes it possible to extract a larger variety of phrases that express the desired relationship than Bouraoui et al. (2020) can. The authors also generate candidate sentences by paraphrasing. In short, they find that both mined and paraphrasing have their usages, and that combinations of template types, e.g., manually constructed and mined, often perform well.

Poerner et al. (2019) question the conclusion by Petroni et al. (2019) that BERT contains factual knowledge derived from the training data. The authors believe that in may cases, BERT simply exploits superficial similarities and general patterns to guess what is most likely. For example, from the fact that a person has a typically French surname, BERT could guess that that person is actually French without having learned the nationality of the particular person. To expose this weakness, (Poerner et al., 2019) remove what they believe are easily guessed pairs of subjects and objects from the data set of (Petroni et al., 2019). They also provide a modified version of BERT, E-BERT, in which the embeddings of entities mentioned in Wikipedia have been replaced by a symbolic entity embedding. They find that E-BERT outperforms both BERT and ERNIE on the trimmed data set, but also that a combination E-BERT and BERT (taking the average of or concatenating the embeddings) give higher accuracy than either on its own.

Rosenbloom (2010) model different types of declarative and procedural memory with what is essentially weighted hypergraphs, in which nodes correspond to actions and conditions, and edges to activation functions. Procedural and declarative memory are distinguished based on the direction in which values are propagated through the hypergraph. The analogy to human cognition is that procedural memory contains information about how to do something, whereas declarative memory concerns facts and events.

3 The role of semantic relations

As the brief account given in the previous section shows, there is a solid body of work on the extraction of relations from language models (see Section 2), to derive facts such as that the birth place of Olga Tokarczuk is Sulechów, Poland, and that the capital of Bolivia is La Paz. Looking to knowledge bases, it is natural to view them as graphs, where nodes represent objects and properties, and edges represent semantic relations. Finally, for

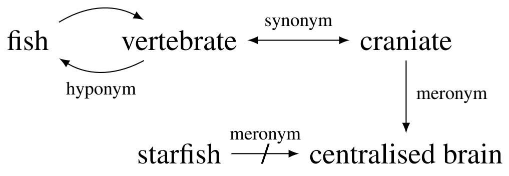

Figure 1: In this work we focus on recovering synonyms, hypernyms, hyponyms, and meronyms from natural language models via probing to understand the prerequisites of integration with knowledge bases.

Synonymy		**Hyponymy**		**Meronymy**	
band	set circle ring	assumption	theory miracle audacity	house	library attic porch
office	agency bureau authority	copper	metal penny policeman	road	bend crossing turnout

Table 1: Instances of the relations synonymy, hypernymy, and meronymy extracted from WordNet.

logical inference, basic semantic relations such as synonymy, hyponymy, hypernymy, and meronymy play a central role. We recall that words are synonyms if they have (nearly) the same meaning; that a hypernym of a concept is a generalisation of that concept (e.g., 'bird' is a hypernym of 'sparrow'), while a hyponym is an instance of the concept (e.g., 'spider' is a hyponym of 'arachnid'), and that a meronym of a concept is a part of the whole (e.g., 'branch' is a meronym of 'tree'); see Table 1 for examples found in WordNet (Fellbaum, 1998).

For logical inference, we can infer that starfish are not fish from knowing that 'heart' is a meronym of 'craniate' but not of 'starfish' (all craniates have hearts whereas starfish do not), 'vertebrate' is a hypernym of 'fish' (fish are vertebrates), and 'craniate' is a synonym of 'vertebrate'. See Figure 1 and Table 1 for further examples.

To achieve a seamless integration of a neural network with a knowledge base of relations and an inference engine, we propose to devise methods for (i) enabling the network to utilise the knowledge base, but fall back on the less certain information in the embedding when necessary and (ii) taking the relations in the knowledge base into account during network training, so that the trained network reflects the contents of the knowledge base. In this endeavour, we believe that particular emphasis should be placed on the treatment of lexico-semantic relations such as meronymy, hyponymy, and synonymy because of their central role in logical deduction and lexical semantics.

4 Empirical study: method

To gain some initial insight into how well state-of-the-art pretrained contextual embeddings handle lexico-semantic relations, we conducted experiments on word embeddings generated by ALBERT (Lan et al., 2020), ROBERTa (Liu et al., 2019), BERT (Wolf et al., 2019), and GPT-2 (Radford et al., 2019). We also included Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) models in our experiments, for comparison. These are all self-supervised learning algorithms, based on neural networks and built to translate words into vector representations. BERT and GPT-2 are transformer models, each having 12 encoder layers. ROBERTa is a retraining of BERT on a larger data set, while ALBERT is an extension of BERT that has a higher data throughput with 10x fewer parameters, and thus scales better.

In contrast to the works discussed in Section 2, we do not extract relations from the embeddings by means of linguistic templates. Rather, we view sentence extraction as an instance of *probing* (Rogers et al., 2018; Conneau et al., 2018; Yaghoobzadeh et al., 2019; Hupkes et al., 2020), a diagnostic method to reveal what aspects of the input the embedding actually encodes. Probing tasks should ideally be agnostic as to the underlying encoder architecture, so that results are transferable between embeddings (Hewitt and Liang, 2019; Dahlgren et al., 2021). Random control tasks (Hewitt and Liang, 2019) are implemented, see discussion in Section 5. In our experiments, we considered the following probing task: Given a pair of word vectors, we ask whether the encoded words are in relation $\mathcal{R}$. This avoids the optimisation problem linked to the choice of template seen in (Petroni et al., 2019).

All experiments are on the English language, and the data set used in our experiments was obtained from WordNet as follows. We first built a vocabulary V by taking the 5 000 most common nouns in the Brown corpus (Kucera and Francis, 1967) and removing those not found in WordNet (Fellbaum, 1998). This resulted in a vocabulary of 3497 words. For each word w in the vocabulary V and target relation $\mathcal{R} \in \{hypernym, meronym, synonym\}$ we then picked words v and v in V such that

Embedding	Synonyms (50.1)	Meronyms (54.2)	Hypernyms (51.0)	Hyponyms (50.7)
Word2Vec	61.5 (1.8)	68.8 (5.0)	69.1 (1.5)	54.1 (1.7)
GloVe	63.2 (2.3)	73.3 (6.0)	68.7 (2.0)	55.7 (1.7)
ALBERT	51.9 (2.6)	48.7 (2.2)	51.2 (1.8)	51.7 (2.9)
ROBERTa	61.7 (1.9)	62.7 (5.9)	64.1 (1.2)	58.2 (2.8)
BERT	56.7 (1.2)	57.2 (3.6)	64.2 (1.6)	51.1 (0.3)
GPT-2	58.0 (1.2)	61.8 (5.3)	65.0 (1.3)	52.4 (2.5)

Table 2: The probing accuracy on the semantic relations, with variance given in parentheses. The accuracy of a "largest class" strategy is shown next to each relation. All transformers give embeddings of 768 dimensions, with word2vec and GloVe using 300 dimension. Each relation contain 1712, 306, 2740, and 1630 samples, respectively.

$(w, v) \in \mathcal{R}$ and $(w, v') \notin \mathcal{R}$, and stored these as triples (w, v, v').

We formulate a classification task for each relation $\mathcal{R}$, and probe each of the investigated models for their ability to capture each relation in their respective embeddings. Each classification task is based on 1 712, 306, 2 740, 1 630 samples for synonyms, meronyms, hypernyms, and hyponyms respectively. We use a linear classifier probe as these better reflect the availablity of the information probed for, as shown in (Hewitt and Liang, 2019; Dahlgren et al., 2021). From (w, v, v'), positive (w, v) and negative (w, v') examples are drawn with equal probability, labeled either 0 or 1, to represent if the tuple represents a negative or a positive pair. The binary labels are given together with either (w, v) or (w, v') as input to the probe by concatenating both word embeddings. We train the probe for 10 epochs using 5-fold cross validation, using softmax activation, dropout of 0.2 to prevent memorising samples, and cross-entropy loss with the Adam optimizer using a $lr = 0.001$. We average the results over 5 runs. The experiment is implemented with Pytorch for CPU and uses the Huggingface (Wolf et al., 2019) library for all pretrained transformers, and the Gensim (Rehurek and Sojka, 2011) library for word2vec and GloVe. The experiments completed within 1 hour on an Intel i7-based Linux laptop with 32GB RAM. The code is available on Github[1].

5 Results and discussion

Table 2 displays the numerical results, with the header row showing, for each relation $\mathcal{R}$, the size of the larger of the two classes. This number coincides with the control tasks implemented to measure selectivity, which are omitted to limit redundancy. The table shows linear probe classification accuracy for each language model, with the variance written out within parentheses. As can be expected, the variance is highest for meronyms where there is least data.

Various observations can be made by comparing the results for the individual embeddings. Particularly worthwhile noting is the fact that GloVe and word2vec performs on par or better than the contextual embeddings, except for the case of hyponyms. This behaviour was seen with 5 and 20 training epochs as well.

The relatively strong performance of the pre-transformer solutions may not be surprising as far as synonyms are concerned, since their construction builds around aligning words found in the same context. However, we would not have expected similar results for hypernyms and even lesser so for meronyms. We note that ALBERT does not accessibly encode any of the relations, resulting in random guesses. This could be because ALBERT is trained using tenfold fewer parameters to produce much smaller embeddings, and might have less room for this type of information. Since ALBERT is comparable in performance to, e.g., BERT on many data sets and other metrics, this needs further investigation to see to what extent these relations are present in the data sets. The complexity of the probe could also be the culprit, as an embedding with lower dimensionality poses a more difficult task for a probe with limited capabilities of separating intertwined concepts. These results do not mirror those of Lan et al. (2020), which indicates that the relations studied here could receive more attention in future evaluations of language embeddings. ROBERTa seems to generally outper-

[1] https://github.com/dali-does/semprof

form the other transformers, especially on the hyponyms, taking into account that not all results are statistically significant. Hypo-/hypernym relations usually follows a tree hierarchy, with hypernyms directed towards the root. This gives a decreasing number of hypernyms, for example, `fish` has six hypernyms but 39 hyponyms in WordNet, and it is likely that less common words will be chosen as a positive example for hyponyms. Weighting the words according to frequency could show different results, but filtering words based on the data the models are trained on is counterproductive to the purpose of these probes. ROBERTa is better able to capture synonyms, which could be an effect of the much larger dataset used in training compared to the other BERT-models leading to more of the less common examples of hyponyms being seen more. One hypothesis on why GPT-2 also shows poor performance is that Wikipedia is removed from the training data. The proposition is that many Wikipedia articles explicitly outlines hyponym relations, e.g. in "*The **cat** is a [domestic species of small carnivorous] **mammal***" [2].

Summarising the results, the fact remains that according to our probes no model covers the relations reliably. If this observation is confirmed by further experiments, it supports the case for a combination of neural networks, traditional relational knowledge bases, and inference engines. With this architecture, established facts could be retrieved from the knowledge base and complemented by less certain facts deduced by the network to cover up for missing information without causing inconsistencies. The results also indicate that a significant threshold should be applied for transferring relational knowledge derived from an embedding to a knowledge base, if this should be done at all, to avoid large error propagation. This is especially important if the "facts" in the knowledge base are considered to be absolute truths rather than tentative findings.

In conclusion, the reliability of the probe could improve with evaluation sets from relations found in knowledge bases, and a correlational study between probing accuracy and downstream NLP tasks could further support the usefulness of studying these relations.

[2] `https://en.wikipedia.org/wiki/Cat`

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT 2021, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7456–7463. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 2126–2136.

Adam Dahlgren, Johanna Björklund, and Frank Drewes. 2021. Perception, memory, and inference: The trinity of machine learning. Digital proceedings of the IJCAI 2021 Workshop NSNLI.

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.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Patrick Hohenecker and Thomas Lukasiewicz. 2020. Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research (JAIR)*, 68:503–540.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

H. Kucera and W. N. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Judea Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Paul S Rosenbloom. 2010. Combining procedural and declarative knowledge in a graphical architecture. In *Proceedings of the 10th International Conference on Cognitive Modeling*, pages 205–210.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark. Association for Computational Linguistics.

Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. 2021. Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases. arXiv:2102.11965.

Po-Wei Wang, Priya Donti, Bryan Wilder, and Zico Kolter. 2019. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6545–6554. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Yadollah Yaghoobzadeh, Katharina Kann, T. J. Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753, Florence, Italy. Association for Computational Linguistics.