Journal of Logic, Language and Information (2024) 33:1–7
https://doi.org/10.1007/s10849-023-09408-y

Special Issue of Natural Logic Meets Machine Learning (NALOMA): Selected Papers from the First Three Workshops of NALOMA

Aikaterini-Lida Kalouli[1] · Lasha Abzianidze[2] · Stergios Chatzikyriakidis[3]

Published online: 12 December 2023

In recent years, there has been a surge of interest in tasks targeting Natural Language Understanding (NLU) and Reasoning. Most recently, Large Language Models (LLMs) such as ChatGPT[1] have received immense attention and have made such NLU tasks seem more tangible as ever. For the creation of these models many research efforts have focused on the creation of massive datasets and the training of huge, deep models reaching human performance, cf. ChatGPT,[2] PALM-2,[3] LLaMMA,[4] Falcon[5] but also Liu et al. (2019), Pilault et al. (2020). The world knowledge encapsulated in such models and their robust nature enables them to deal with diverse and large amounts of data in an efficient way. However, it has been repeatedly shown that such models fail to solve basic human inferences and lack generalization power. When presented with differently biased data (Poliak et al., 2018; Gururangan et al., 2018; Kalouli et al., 2023) and smaller datasets with fewer or less diverse phenomena (Bender et al., 2021), or with inferences containing hard linguistic phenomena, (Dasgupta et al. 2018; Nie et al. 2018; Naik et al. 2018; Glockner et al. 2018; Richardson et al. 2020; McCoy et al. 2019; Bernardy amb Chatzikyriakidis 2019; Yanaka et al. 2020, to name only a few), they struggle to reach the baseline. Explicitly detecting and solving these weaknesses is only partly possible, e.g., through appropriate datasets, fine-tuning or appropriate prompting, because such models act like black-boxes with

[1] https://openai.com/blog/chatgpt
[2] https://openai.com/blog/chatgpt
[3] https://blog.google/technology/ai/google-palm-2-ai-large-language-model/
[4] https://ai.meta.com/blog/large-language-model-llama-meta-ai/
[5] https://falconllm.tii.ae/

This work was completed when the first editor was at the CIS in LMU Munich.

✉ Aikaterini-Lida Kalouli
katerina.kalouli@hotmail.com

1 Bundesdruckerei GmbH, Berlin, Germany

2 Utrecht University, Utrecht, The Netherlands

3 University of Crete, Crete, Greece

low explainability. At the same time, another strand of research has continued to target more traditional approaches to reasoning, employing some kind of logic or semantic formalism. Such approaches excel in precision, especially of inferences with hard linguistic phenomena, e.g., negation, quantifiers, modals, etc. (Bernardy and Chatzikyriakidis 2017; Yanaka et al. 2018; Chatzikyriakidis and Bernardy 2019; Hu et al. 2019; Abzianidze 2020, to name only a few). However, they suffer from inadequate world knowledge and lower robustness, making it hard for them to compete with state-of-the-art models. Thus, lately, a third research direction seeks to close the gap between the two approaches by exploring how the strengths of the two approaches can be combined and their weaknesses mitigated, e.g., through hybrid approaches.

Attempts to combine distributional and symbolic representations to tackle NLU tasks have been pursued in three main directions. One strand of research has used linguistic or formal semantic features as additional input to systems that create distributional representations, e.g., Padó and Lapata (2007), Bjerva et al. (2014), Levy and Goldberg (2014), Bowman et al. (2015), Chen et al. (2018). Another strand of research has attempted the opposite: to use distributional features as input to systems that create symbolic representations, e.g., May (2016), van Noord et al. (2018), Oepen et al. (2020). Both these research directions have laid their focus on one of the frameworks and have only used the other one in a complementary manner. The third research direction has attempted to lay an equal focus on the two frameworks by combining symbolic and distributional aspects in the final representation, e.g., Lewis and Steedman (2013), Beltagy et al. (2016), Kalouli et al. (2019), Krishna et al. (2022), marrying traditional reasoning paradigms with neural approaches, e.g., Liang et al. (2017), Ebrahimi et al. (2021) or aiming at explainable Artificial Intelligence (AI) (Calegari et al., 2020). We see such hybrid research efforts as promising not only to overcome the described challenges and advance the field but also to contribute to the symbolic-deep learning “debate" that has emerged in the field of NLU.

Indeed, hybrid approaches have been pursued in several sub-fields of NLU, such as Natural Language Inference (NLI), Question-Answering (QA), Sentiment Analysis and Dialog. Concerning NLI, recent research by Kalouli et al. (2020) proposes a hybrid approach where a trained classifier learns whether the symbolic or the deep learning component of the system should be trusted based on the nature of the pair, i.e., on whether it involves complex linguistic phenomena and thus requires precise reasoning or whether robustness and world-knowledge are necessary. Within medical NLI, Wu et al. (2019) present an approach of an ensemble model, based on one symbolic and two deep learning encoders. The symbolic encoder is a syntax encoder, capturing structural information of the sentences, while the deep learning encoders are responsible for converting the text into distributional representations and injecting domain knowledge into the model. The QA field has attracted similar interest in hybrid methodology. Yi et al. (2018) propose a neural-symbolic visual question-answering system, which first recovers a structural scene representation from the image and a program trace from the question and then executes the program on the scene representation to obtain an answer. Honda and Hagiwara (2019) employ a combination of deep learning models, Neural Machine Translation and Word2Vec training to learn the symbolic processing performed by a Prolog system and use it to build a QA system. Within the field of Sentiment Analysis, Hu et al. (2017) propose a framework that enhances various

types of neural networks (e.g., CNNs and RNNs) with declarative first-order logic rules by transferring the structured information of logic rules into the weights of neural networks. They show that their approach is able to outperform the state-of-the-art in Sentiment Analysis and Named Entity Recognition. More recent research in hybrid Sentiment Analysis has been conducted by Cambria et al. (2020), who implement a new version of SenticNet (Cambria et al., 2018), a knowledge base used for sentiment analysis, by employing a top-down (symbolic) and a bottom-up (subsymbolic) approach. They use logic and semantic networks to encode meaning and deep learning architectures to implicitly learn syntactic patterns from the data. Kalatzis et al. (2016) and Eshghi et al. (2017) combine reinforcement learning with a symbolic dynamic model of syntax (Dynamic Syntax) and demonstrate the effectiveness of such approach in bootstrapping dialog data from very minimal data.

Against this backdrop of hybrid approaches in NLU, we have promoted this research direction and fostered fruitful dialog between the two disciplines by establishing the NALOMA (Natural Logic Meets Machine Learning) workshop series in 2020 [6] and continuing it since then. [7] The workshop, which started out with a focus on NLI, aims to bring together researchers working on hybrid methods in any subfield of NLU, including but not limited to NLI, QA, Sentiment Analysis, Dialog, Machine Translation, Summarization, etc. The workshops have also attracted researchers working on one of the two disciplines but interested in moving into the hybrid direction. Topics that have been part of the workshops include: NLU systems that integrate logic-based/symbolic methods with neural networks, explainable NLU models, opening the "black box" of deep learning in NLU, downstream hybrid NLU applications, comparison and contrast between symbolic and deep learning work on NLU, etc.

With this special issue, we would like to put together extended versions of several selected contributions to the NALOMA series. Specifically, the issue contains the following four contributions.

Assessing the Strengths and Weaknesses of Large Language Models The opening paper by Shalom Lappin includes a clear, calm and insightful discussion of the strengths and weaknesses of Large Language Models (LLMs). It provides a very timely and a welcome contribution to the current, and potentially overhyped discussion on LLMs. The article carefully examines the arguments against the use of LLMs and takes a balanced stance according to which LLMs are far more than stochastic parrots, but at the same time, the question of whether these models have anything to say in the areas of human language learning and linguistic representation has been largely left unanswered.

Capturing the Varieties of Natural Language Inference: A Systematic Survey of Existing Datasets and Two Novel Benchmarks Gubelmann et al. specifically focus on the task of NLI. After providing an extensive survey on the types of inference (from a theoretical point of view) and on the current scenery of neural NLI models and datasets, they lay the foundations to discuss the problem of generalization of these

[6] https://typo.uni-konstanz.de/naloma20

[7] https://typo.uni-konstanz.de/naloma21, https://sites.google.com/view/naloma22, https://sites.google.com/view/naloma4

models based on the theoretical notions of inference. Particularly, they use a dataset of the argumentative writing field to evaluate and criticize the strengths of neural NLI models on such kinds of inferences. In their second experiment, they create their own dataset, which focuses on quantifiers and deductively valid inferences. Again, they probe LLMs on these datasets and make conclusions about their capabilities in this area. Last, the authors explore options for fine-tuning and optimizing the models, also in comparison with a symbolic system of NLI.

Monotonicity Reasoning in the Age of Neural Foundation Models The paper by Chen and Gao presents three methods to tackle monotonicity reasoning using deep learning and large language models. The first approach utilizes a Tree-LSTM with syntactic tree structures and a multi-hop self-attention aggregator to classify natural language inference problems. The second approach represents a pipeline of rule-based and neural components. The NLI pipeline first detects polarities of words based on monotone operators and a sentence structure. Then, the search engine attempts to gradually rewrite a premise into a hypothesis. One of the components in the rewriting search is neural-based which detects paraphrases that are beyond monotonicity calculus. The third approach exploits LLMs, including GPT3.5, to classify monotonicity inference problems in zero- and few-shot learning experiments. The overall conclusion of the paper is three-fold: LLMs are far from mastering monotonicity reasoning, the underlying tree structures do help in classifying monotonicity inferences, and joint reasoning with symbolic and neural components can set state-of-the-art on monotonicity reasoning.

Monotonic Inference with Unscoped Episodic Logical Forms: From Principles to System Kim et al. propose a theoretical framework and its implementation for monotonicity inference with Unscoped Episodic Logical Forms (ULFs), where the latter is an Episodic Logic formula with unresolved scope, anaphora, and word senses. The implemented system is mainly a pipeline of rule-based components. The inference process represents a forward search from the premises to the hypothesis. The authors additionally extend the baseline system in three ways: (1) use lexical information from the hypothesis to better guide the forward inference process, (2) consider multiple possible scopings for a sentence, and (3) base matching of an obtained conclusion and the hypothesis on surface forms to abstract from possible parsing errors introduced in ULFs. Both baseline and extended systems are evaluated on the generalized quantifier section of the FraCaS NLI dataset. The results show that each extension separately improves the baseline but jointly yields an average improvement. This is partially due to (2) and (3) introducing wrong entailment relations.

References

Abzianidze, L. (2020). Learning as abduction: Trainable natural logic theorem prover for natural language inference. In Proceedings of the 9th joint conference on lexical and computational semantics (pp. 20–31). Association for Computational Linguistics, Barcelona, Spain

Beltagy, I., Roller, S., Cheng, P., Erk, K., & Mooney, R. J. (2016). Representing meaning with a combination of logical and distributional models. *Computational Linguistics, 42*(4), 763–808.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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 '21 (pp. 610–623). Association for Computing Machinery, New York, NY, USA

Bernardy, J.-P., & Chatzikyriakidis, S. (2017). A type-theoretical system for the FraCaS test suite: Grammatical framework meets Coq. In Proceedings of the 12th international conference on computational semantics (IWCS)–long papers, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Montpellier, France.

Bernardy, J.-P., & Chatzikyriakidis, S. (2019). What kind of natural language inference are NLP systems learning: Is this enough? In ICAART (2) (pp. 919–931).

Bjerva, J., Bos, J., van der Goot, R., & Nissim, M. (2014). The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), (pp. 642–646). Association for Computational Linguistics, Dublin, Ireland.

Bowman, S. R., Potts, C., & Manning, C. D. (2015). Recursive neural networks can learn logical semantics. In Proceedings of the 3rd workshop on continuous vector space models and their compositionality (pp. 12–21). Association for Computational Linguistics, Beijing, China.

Calegari, R., Ciatto, G., & Omicini, A. (2020). On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale, 14*, 7–32.

Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In AAAI.

Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In Proceedings of the 29th ACM international conference on information & knowledge management, CIKM '20 (pp. 105–114). Association for Computing Machinery, New York, NY, USA.

Chatzikyriakidis, S., & Bernardy, J. -P. (2019). A wide-coverage symbolic natural language inference system. In Proceedings of the 22nd nordic conference on computational linguistics (pp. 298–303). Linköping University Electronic Press, Turku, Finland.

Chen, Q., Zhu, X., Ling, Z. -H., Inkpen, D., & Wei, S. (2018). Neural natural language inference models enhanced with external knowledge. In Proceedings of the 56th annual meeting of the association for computational linguistics (vol. 1, pp. 2406–2417). Association for Computational Linguistics, Melbourne, Australia.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., & Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings. CoRR arXiv:1802.04302

Ebrahimi, M., Eberhart, A., Bianchi, F., & Hitzler, P. (2021). Towards bridging the neuro-symbolic gap: Deep deductive reasoners. *Applied Intelligence, 51*, 6326–6348.

Eshghi, A., Shalyminov, I., & Lemon, O. (2017). Bootstrapping incremental dialogue systems from minimal data: The generalisation power of dialogue grammars. In Proceedings of the 2017 conference on empirical methods in natural language processing (pp. 2220–2230). Association for Computational Linguistics, Copenhagen, Denmark.

Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. In Proceedings of the 56th annual meeting of the association for computational linguistics (vol. 2, pp. 650–655). Association for Computational Linguistics.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. In Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies (vol. 2, pp. 107–112). Association for Computational Linguistics.

Honda, H., & Hagiwara, M. (2019). Question answering systems with deep learning-based symbolic processing. *IEEE Access, 7*, 152368–152378.

Hu, H., Chen, Q., & Moss, L. (2019). Natural language inference with monotonicity. In Proceedings of the 13th international conference on computational semantics– short papers (pp. 8–15). Association for Computational Linguistics, Gothenburg, Sweden.

Hu, Z., Ma, X., Liu, Z., Hovy, E., & Xing, E. (2017). Harnessing deep neural networks with logic rules. In Proceedings of the 54th annual meeting of the association for computational linguistics (vol. 1, pp. 2410–2420). Association for Computational Linguistics, Berlin, Germany.

Kalatzis, D., Eshghi, A., & Lemon, O. (2016). Bootstrapping incremental dialogue systems: Using linguistic knowledge to learn from minimal data. In Proceedings of the NIPS 2016 workshop on learning methods for dialogue.

Kalouli, A. -L., Crouch, R., & de Paiva, V. (2020). Hy-NLI: A hybrid system for natural language inference. In Proceedings of the 28th international conference on computational linguistics (pp. 5235–5249). International Committee on Computational Linguistics, Barcelona, Spain (Online).

Kalouli, A. -L., Crouch, R., & dePaiva, V. (2019). GKR: Bridging the gap between symbolic/structural and distributional meaning representations. In Proceedings of the first international workshop on designing meaning representations (pp. 44–55), Association for Computational Linguistics, Florence, Italy.

Kalouli, A.-L., Hu, H., Webb, A. F., Moss, L. S., & de Paiva, V. (2023). Curing the SICK and other NLI maladies. *Computational Linguistics, 49*, 1–45.

Krishna, A., Riedel, S., & Vlachos, A. (2022). ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics, 10*, 1013–1030.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In Proceedings of the 52nd annual meeting of the association for computational linguistics (vol. 2, pp. 302–308). Association for Computational Linguistics, Baltimore, Maryland.

Lewis, M., & Steedman, M. (2013). Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics, 1*, 179–192.

Liang, C., Berant, J., Le, Q., Forbus, K. D., & Lao, N. (2017). Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. In Proceedings of the 55th annual meeting of the association for computational linguistics (vol. 1, pp. 23–33). Association for Computational Linguistics, Vancouver, Canada.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 4487–4496). Association for Computational Linguistics, Florence, Italy.

May, J. (2016). SemEval-2016 task 8: Meaning representation parsing. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016) (pp. 1063–1073). Association for Computational Linguistics, San Diego, California.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 3428–3448). Association for Computational Linguistics, Florence, Italy.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., & Neubig, G. (2018). Stress test evaluation for natural language inference. In Proceedings of the 27th international conference on computational linguistics (pp. 2340–2353). Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Nie, Y., Wang, Y., & Bansal, M. (2018). Analyzing compositionality-sensitivity of NLI models. CoRR arXiv:1811.07033

Oepen, S., Abend, O., Abzianidze, L., Bos, J., Hajic, J., Hershcovich, D., Li, B., O'Gorman, T., Xue, N., & Zeman, D. (2020). MRP 2020: The 2nd shared task on cross-framework and cross-lingual meaning representation parsing. In Proceedings of the CoNLL 2020 shared task: Cross-framework meaning representation parsing (pp. 1–22). Association for Computational Linguistics, Online.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics, 33*(2), 161–199.

Pilault, J., Elhattami, A., & Pal, C. (2020). Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameter & less data.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R. & Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In Proceedings of the 7th joint conference on lexical and computational semantics (pp. 180–191). Association for Computational Linguistics, New Orleans, Louisiana.

Richardson, K., Hu, H., Moss, L. S., & Sabharwal, A. (2020). Probing natural language inference models through semantic fragments. In The 34th AAAI conference on artificial intelligence, AAAI 2020, the 32nd innovative applications of artificial intelligence conference, IAAI 2020, the 10th AAAI symposium on educational advances in artificial intelligence, EAAI 2020 (pp. 8713–8721). AAAI Press, New York, NY, USA.

van Noord, R., Abzianidze, L., Toral, A., & Bos, J. (2018). Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics, 6*, 619–633.

Wu, Z., Song, Y., Huang, S., Tian, Y., & Xia, F. (2019). WTMED at MEDIQA 2019: A hybrid approach to biomedical natural language inference. In Proceedings of the 18th BioNLP workshop and shared task (pp. 415–426). Association for Computational Linguistics, Florence, Italy.

Yanaka, H., Mineshima, K., Bekki, D., & Inui, K. (2020). Do neural models learn systematicity of monotonicity inference in natural language? In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 6105–6117). Association for Computational Linguistics, Online.

Yanaka, H., Mineshima, K., Martínez-Gómez, P., & Bekki, D. (2018). Acquisition of phrase correspondences using natural deduction proofs. In Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies (vol. 1, pp. 756–766). Association for Computational Linguistics, New Orleans, Louisiana.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. (2018). Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems.* (Vol. 31). Curran Associates Inc.