A Review on Neuro-symbolic AI Improvements to Natural Language Processing

M. Keber*†, I. Grubišić*†, A. Barešić*, A. Jović†
* Ruđer Bošković Institute, Zagreb, Croatia
† University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
miha.keber@irb.hr

***Abstract*—Symbolic artificial intelligence (AI) reflects the domain knowledge of experts and adheres to the logic of the subject area, rules, or any relations between entities. Connectionist (neuro) approaches based on artificial neural networks are excellent for extracting abstract features, contextualizing, and embedding interactions between features. When connectionist and symbolic approaches are properly aligned in a model, they benefit from complementary strengths; the combination is referred to as a hybrid or neuro-symbolic artificial intelligence (NSAI) model. The advantages that NSAI brings to the field of natural language processing (NLP) have received little attention from researchers in recent years. Therefore, in this review, we focus on the impact of neuro-symbolic approaches for NLP tasks, i.e. text classification, information extraction, machine translation, and language understanding. Relevant research articles from Scopus, Web of Science, and Google Scholar were carefully examined using appropriate keywords in the period from 2019 to 2024. The review aims to show the types of NSAI systems, identify the motivation for using NSAI, evaluate the use of additional annotations for content description, and briefly describe how the neuro-symbolic connection improves the methodology and enables trustworthy and explainable AI systems in current NLP research. The review also highlights areas of application and improvements achieved by NSAI approaches in benchmarks.**

***Keywords*—neuro-symbolic artificial intelligence, natural language processing, knowledge representation, deep learning**

I. Introduction

Existing knowledge is often represented using symbols and rules, both readable by computers and largely understandable by humans. This kind of representation makes symbolic artificial intelligence (AI) interpretable and explainable to users. Symbolic models usually form a knowledge base (KB) and include ontologies, sets of rules, and domain logic. They are created using the knowledge of domain experts and require considerable effort. With high-dimensional raw data, it is difficult to extract relevant rules that are robust to noise and take into account interactions between attributes. Unlike symbolic models, deep learning (DL) models learn directly from high-dimensional data, by automatically extracting features from data. Unfortunately, DL models are mostly black-box models that can hardly be interpreted. Integrating symbolic and DL models, neuro-symbolic artificial intelligence (NSAI) leverages the supporting strengths of both disciplines to compensate for their imperfections.

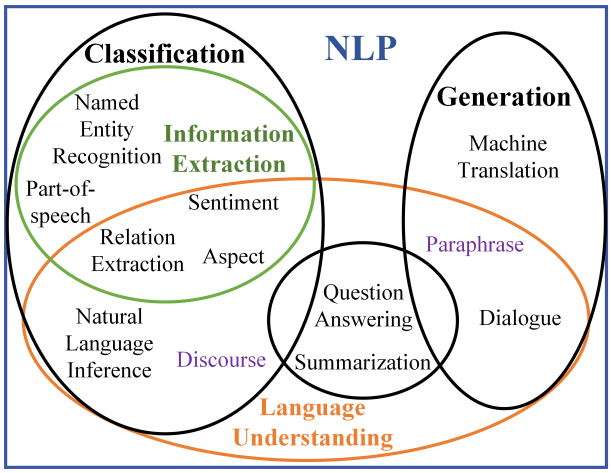

Fig. 1: Categorization of common tasks in NLP, tasks in black color are within the scope of this survey.

Neural language models (NLMs) based on DL represent tokens in a continuous vector space and were first introduced by Mikolov et al. [1] in 2013. They proposed learning unique word representations from a corpus while retaining syntactic and semantic linguistic information; this process is called pre-training. Since the meaning of the word changes depending on the neighboring words, i.e. the context of the text, newer NLMs such as ELMo [2] or BERT [3] create context-dependent representations. Larger and deeper pre-trained NLMs, also called large language models (LLMs), contain representations with high-quality features and exhibit emerging properties [4], which is the ability of a model to learn specific knowledge without specific training. Consequently, both scientists and industry drivers increased the number of parameters in NLMs, exceeding trillions of them [5]. However, pre-training, domain adaptation, and fine-tuning of LLMs require resources not available to most researchers.

Motivated by the above and with the high cost of annotating data for building high-quality models, this work aims to review NSAI methods for solving the following NLP tasks: 1) named entity recognition (NER); 2) relation extraction (RE); 3) question answering (QA); 4) natural language inference (NLI); 5) sentiment analysis; 6) machine translation (MT); and 7) summarization (summ), as depicted in Figure 1. The papers reviewed in this study were published between 2019 and 2024 and were carefully

selected based on sets of keywords[1] from three sources: Scopus, Web of Science, and Google Scholar, further filtered for relevance.

We contribute by comparing the performance of each NSAI method to the corresponding best-performing model from an online resource[2], although the score may not necessarily be the intended goal of the surveyed NSAI method. Another contribution is in summarizing the NSAI applications to NLP tasks from the past 5 years and aggregating them to a higher level of abstraction in a following manner. We researched where knowledge was introduced to a DL model (its regularization type), and the effects of that fusion by several categories: data efficiency (DE), performance, explainable AI (XAI), and parameter efficiency (PE). We highlight the key point achieved by an introduced NSAI method, and how it was developed. In this way, we give the reader an objective insight into the effort-reward ratio of NSAI for NLP.

In Section II, the existing surveys regarding the use of NSAI in NLP are listed and briefly described. Section III provides fundamental information about NLP, symbolic AI, and DL. Section IV contains the main insights of the survey, presented in a condensed form within Table I. Section V concludes the paper.

II. Related Work

The usual focus of related work is on incorporating symbolic domain knowledge into existing NLP models. The classification of events in the biomedical domain by Frisoni et al. [6] described knowledge-based informed (KI) DL architectures. The study described datasets, KBs, and machine learning models for event extraction (EE) with an evaluation of performance. In contrast, the work of Oltramari et al. [7] focused more on several knowledge injection techniques for NLMs with a description of the KBs for question-answering (QA) tasks. Yang et al. [8] distilled the coupling methods for knowledge representation with pre-trained NLMs using the attention mechanism [9]. Liu et al. [10] discussed several NSAI methods optimized using reinforcement learning for weakly supervised information extraction. Hamilton et al. [11] described neuro-symbolic models and their constituting parts and evaluated different effects achieved without quantification. In addition, they provided summary statistics on trends in industrial applications and the quality of research studies [11]. When compared to [8], [10], [11], we simplify the overview by extracting the methodology for coupling symbolic and connectionist methods, and focus on evaluating the performance and different machine learning effects achieved on text data[3].

III. Fundamentals

Creating pragmatic NLP task-specific hybrid models requires an understanding of knowledge representation (KR) and representation learning. Therefore, this section covers introductory methods and notions related to NLP tasks, as well as symbolic and connectionist approaches related to NLP.

A. Natural Language Processing Tasks

One of the oldest NLP tasks began in the 1940s when a generative sequence-to-sequence mechanical translation relying on syntactic features was pursued. Today, machine translation (MT) still benefits from part-of-speech (POS) tagging and parsing, where each word is tagged and the structure of a sentence is created resembling different linguistic levels: lexical, dependency, and syntactic. The field of information extraction (IE) gained popularity after 1995, when Grishman and Sundhein [12] presented the first NER challenges still burdened by inter-annotator agreement and what the true gold standard labels are. Presently, many NLP tasks benefit from specialized public KBs, e.g., word polarity SenticNet [13], which specializes in sentiment analysis, emotion detection, and stance detection to reduce the influence of negations, sarcasm, and cynicism occurring in texts.

Natural language inference (NLI) evaluates model language understanding of the semantic and syntactic structure by having to classify the hypothesis concerning the premise: entailment, contradiction, or neutral. Predictors using DL for NLI tasks suffer from word co-occurrence between premise and hypothesis, predetermined by DL use of shortcuts. Gaining momentum lately, QA is used to evaluate LLM stored knowledge and can be set up as a classification or generation task, where the aim is to select the correct label from a multiple-choice answer or to generate an answer [14].

B. Symbolic Knowledge Representations

A symbolic AI system usually consists of a knowledge base (KB), which contains symbolic terms and data, and an execution engine. The engine reasons about data, where a model satisfies a KB if it is consistent for all KB axioms. KBs may consist of common knowledge, e.g. the CYC project [15], or linguistic domain-specific knowledge as in WordNet [16], or biological abnormalities in human disease knowledge as in Human phenotype ontology (HPO) [17]. All these KBs are often used in NLP. Ontologies, as a specific form of KB, are created with a predefined formalization of the relations between concepts, namely axioms of a description logic.

First-order logic (FOL) defines a logical language as a tuple of predicates, functions, constraints, and

[1]Neuro-symbolic: nsai, neuro-symbolic, neurosymbolic, knowledge-injection, knowledge informed, prior knowledge, knowledge injection, neural-symbolic and connectionist. NLP: natural language, nlp, language processing, language mode*, machine translation, information extraction, named entity, text classification, language understanding, aspect mining, topic model* and summarization. Symbolic: symbolic, logic, predicate, fol, syntax, morphology, pos, tagging, ontology and rdf.

[2]The best-performing models were retrieved on January 31 2024 from https://paperswithcode.com/ leaderboard platform.

[3]We did not evaluate visual question answering. Instead, only different modalities and views of textual data were considered.

variables. The FOL formulas can specify 1) simple axioms, e.g., if an entity is a person, then it is not a number: $\forall x, Person(x) \wedge \neg Number(x)$, unless we have a person named "One", or 2) general rules like $\forall Gene \exists Locus(partOf(Gene, Locus) \wedge partOf(Locus, Chromosome))$. NLP usually leans towards fuzzy logic by allowing multiple semantics of words and propositions knowing only the probability of being *True* or *False*, thus allowing ambiguous meanings. NSAI researchers often resort to the probabilistic soft logic (PSL) models to adapt probabilistic graphical models (PGMs).

C. Deep Learning in NLP

Recurrent neural networks (RNNs) extract features from arbitrary sequence lengths. The long short-term memory (LSTM) and gated recurrent unit (GRU) RNN cells were introduced to avoid the gradient issues and create high-capacity expressive RNNs, which eventually evolved into bidirectional LSTMs (bi-LSTM) [18] and GRUs (bi-GRU). In 2014, Bahdanu et al. [9] developed an encoder-decoder MT model contextualizing input representations to the decoder as attention-weighted aggregation of encoder outputs. A generalization of the former multi-head attention (MHA) mechanism was introduced by Vaswani et al. [19] in a transformer architecture. MHA creates an attention matrix that can be interpreted as a weighted, fully connected graph adjacency matrix to induce ordering upon a graph-like structure. The inputs require a positional encoding, an absolute encoding [19], and a later-suggested relative positional encoding within the MHA [20]–[22]. Graph attention networks (GAT) [23] use an attention mechanism between connected nodes and the graph transformer network [24].

IV. Improving NLP with NSAI

This section delves into efficient NSAI methods that overcome distinct limitations of DL models in NLP tasks.

In the examined work, we identified three general paths for improving NLP with NSAI that are elaborated in subsection IV-A. The first path focuses on creating rules that constrain the model, i.e. create a new KB for key variables or features. The second path includes methods focused on dataset preprocessing – establishing links to existing KB terms. This includes the modifications to DL architecture and data aggregations. The third path includes methods based on incorporating the symbolic engine in NLM prompt pipelines for some predefined features. All these paths in NSAI leverage KBs to regularize the DL model, accounting for the established domain expert understanding and often necessary additional heuristics. Some of the additional heuristics include the number of neighbors to consider, and choosing a subset of relations from a KB.

We compare the performance of the incorporated NSAI method to the baseline state-of-the-art method, wherever the performance data are available. We also categorize and evaluate the effects that the NSAI method inclusion brings in terms of DE, performance, XAI, and PE. Lastly, we direct attention toward assessing the societal advantages (human-centric AI) from the latest research, emphasizing the workings of NSAI models in practice.

A. Symbolic KR for DL

In the `Regularization` column of Table I, we summarize four types of regularization used to incorporate symbolic KR in a DL model: 1) the model was constrained with rules; 2) a modification (mod) of the DL architecture; or 3) the input attributes or hidden representations were aggregated (agg.) with symbolic KR; or 4) pipeline.

1) Rule-Constrained Model: The first methodology on the rule-constrained model by Nandwani et al. introduced a set of rules as constraints in the Lagrangian dual [25]. Other methods were realized based on teacher-student methodology, distilling symbolic knowledge to a DL model. One approach by Chen et al. used expectation maximization (EM) optimization to distill PSL model knowledge into an RNN [26] for weakly supervised learning, thus bringing about the best scoring model only somewhat weaker than the inter-annotator gold standard. The approach rules out impossible outcomes, e.g. if a token is labeled as *inside-organization*, then its previous token must be either *inside-organization* or *begin-organization*. This approach proposed by Aakur and Sarkar focuses on reducing the fraction of labeled data required for training a common-sense NLI model into a new task [27]. They proposed a symbolic teacher from pattern theory using the ConceptNet KB, which chooses the lowest energy conceptualizations for NLI and QA based on semi-supervised learning. The performance of rule-constrained models was checked using NER [25], [26], [28], and all models were aimed to improve data efficiency (DE).

2) Modifications to DL Architecture: Demonstrating that linguistic features are still a valuable resource for MT, a constituency tree was applied by Nguyen et al. to modify the MHA block into tree-structured attention [29] for the transformer. Zhang et al. [30] identified redundant heads of MHA and then added constituency tree information to the local-phrasal position matrix, enabling better reflection of syntactic relations between elements. On the other hand, using Stanford's CoreNLP, Li et al. [31] extract factual relations and impose relations in the adjacency matrix in a parallel branch of MHA. A factual relation mask matrix was constructed by factual relation tuples obtained from CoreNLP and was used in the encoding procedure. Most MHA modifications were successful in enhancing the performance effect of NSAI in MT [29]–[31]. In the classification NLI and QA in medicine, Kang et al. [32] first used NER and RE to connect entities and relations with a medical evidence dependency KB. With these dependency graphs, the authors modified the MHA adjacency matrix of a pre-trained BioBERT. At the time Kang et al. [32] provided best-performing model for PubMedQA, yet two years later a significant performance increase was achieved by Chen et al. [33]. Liang et al. used

Sentic GCN-BERT to construct an adjacency matrix from the dependency tree with weights and a mask determined by the SenticNet ontology to improve overall sentiment prediction performance [34].

3) Aggregations: Input relations refer to the injection of knowledge into a DL model by modifying its input. With dependency POS tags, Chen et al. [35] generated rules that are used as input to an LSTM cell. Using domain-specific ontologies to create multi-modal (MM) free-text representations with shortest dependency paths (SDP) from ontologies resulted in an increased F1 for the RE performance [36]. However, Wu et al. [37] recommended using representations from different language models for DE sentiment and NLI classification [37]. They used BERT, ConceptNet-Numberbatch, and GloVe, where the adjacency matrix was created from ConceptNet ontology relations, and the final representation was optimized with a GAT. Karpov et al. [38] aggregated data from a social network graph using deep walk (DW), matrix factorization (SVD), and a feed-forward network, once for an additional input representation and a second time for aggregation of a parallelized layer in BERT by generated weights for a quicker convergence of perplexity.

4) Pipelines: Recent models used augmented prompts to make API calls to obtain valuable proof-tested calculations within the prompt pipeline [39], [40]. These models were shown to be parameter-efficient (PE). With the availability of LLMs, Tan et al. [41] created aspect-based summaries using ConceptNet and WikipediaDB using weakly-supervised learning. Using aspects and term frequency - inverse document frequency (TF-IDF) for word ranking from aspects, relevant Wikipedia KB prompt tuning using the BART model [42] increased performance when trained on 0.4% of labeled data [41]. Without tuning the model Han et al. [43] used rules to inject a prefix that dynamically influences the MHA into a prompt for each shot at NER and RE. Showing validity of adding context to each separate input, compared to DL end-to-end trained model by Li et al. [44], exhibited lower F1 score on OntoNotes5.0 dataset. Another prompt-informing approach with RoBERTa and concepts from ConceptGraph for stance detection achieved a top score [45].

B. Effect Categories

Effect categories are focused on providing a list of tangible benefits that NLP credits from NSAI. The following subsections show why to use NSAI, while Table I reveals how and for which NLP task NSAI was used.

1) Data Efficiency: DE refers to the ability of an NSAI method to achieve better performance than the DL counterpart when training the models on a fraction of training data. For example, NSAI models for NER benefit from PSL [25], [35], significantly enhancing model performance when using only 1% of training data to update the weights. BERT-type DE models [27], [37], [41] enhanced scores by more than 20% when trained on less than 1% of labeled data. Still, high efficiency and great results were obtained with RNN architectures [25], [35] when less than 5% of labeled data was used. A somewhat different DE-focused method described in [26] used multiple annotator noisy labels for NER and sentiment analysis with an estimation of annotators' reliability.

2) Performance: NSAI models achieve better performance than pure DL approaches by significantly enhancing the existing DL architectures [29], [30], [32], [35], [36], [38], [45], [46], which is visible in Table I column `DL Model` and `Improvement`. As a part of a hybrid NSAI method, Dai et al. [28] used domain-specific rules to create a KB and populate the cancer registry table. Alternatively, for social dialogue, the MT tree transformer [29] was enhanced even further by adding factual knowledge to the model [31]. It would appear, however, that DL end-to-end cyclic parameter sharing showed the most promising results for a transformer-based MT architecture [47].

3) XAI: Explainable AI uses NSAI models to provide explanations through the use of a KB. Biomedical domain EBRO ontology-generated explanations were judged by human-in-the-loop for NER [48], where a person received and evaluated both explanations and predictions from the NER classifier. Also, an XLNet model in [49] used a surrogate model to explain sentiment predictions with POS constituency importance within a Yelp review. Siyaev et al. [50] constructed a specific dataset from Boeing manuals which was used to create an AI system enabling virtual reality educational walk-throughs for maintenance personnel.

4) Parameter Efficiency and External APIs: For instance, the BERT KI model [46] can be competitive to the three times bigger BioMegatron [51] by injecting relevant knowledge from Reactome, SemMedDB, and other biomedical ontologies for NER. Regrettably, most authors did not provide the number of NLM parameters in the studies for us to evaluate PE. The external API refers to calling existing programs to produce facts, namely using a calculator, python interface, or an MT model. As we mentioned before, QA is one of the main tasks to evaluate complex reasoning and LLM memory. To predict on datasets involving mathematics or general knowledge LLMs use external API calls to receive an exact result or additional domain context and often do another inference forward pass through the frozen LLM with the new knowledge [52]. With the use of APIs, the number of necessary parameters using GPT-J (6.7B) and CODEX (14.8B) were efficiently reduced [39], [40]. Yet, some studies, e.g., [40] cannot match the results when compared to a huge LLM [53].

C. Human-centric AI

Unfortunately, LLMs are still lacking in explainability. Contrary to this fact, human-centric AI requires fairness and trustworthiness, which are crucial for high-stakes AI system application [54], [55]. Computer-assisted diagnostics in healthcare, as a longstanding and prospective way to increase quality of life, demands ethical solutions and

TABLE I: SUMMARY OF NSAI METHODS FOR NLP APPLICATIONS. THE TYPE OF NEURO-SYMBOLIC FUSION IS INDICATED IN THE TYPE (KAUTZ) COLUMN, WITH BRIEF ELABORATION IN SYMBOLIC->DL, INDICATING WHERE THE SYMBOLIC MODEL INFLUENCES THE DL MODEL WITH A FURTHER DESCRIPTION OF "->" IN COLUMN REGULARIZATION. COLUMNS VALUES IN: DL MODEL, BENCHMARK DATASET, BASELINE OR STATE-OF-THE-ART AND IMPROVEMENT AND NSAI MODEL PERFORMANCE CORRESPOND ROW-WISE TO EACH OTHER WHERE MULTIPLE CELL ROWS ARE INDICATED BY COMMA "," AND A NEW LINE AFTER A VALUE IN A CELL. WHEN THESE COLUMNS DIFFER IN NUMBER OF ROW-VALUES WITHIN A CELL THE PRIOR STATED VALUE IS CONSIDERED RELEVANT. FINALLY, LEARNING APPROACH AND OPTIMIZATION ARE SPECIFIED.

Model Name	NLP Task	Application Area	Type (Kautz)	Effect Category	Symbolic KR	Symbolic ->DL	Regularization	Learning/ Tuning	Optimization	DL Model	NSAI Model Performance	Benchmark Dataset	Baseline or State-of-the-art	Improvement
2019 [25]	NER + POS SRL	Biomedical	Type 4	DE	M Rules	Rules->LAG Constraints	Constrained with rules	Semi-Sup	Min-max LAG	Bi-LSTM	68.71 F1 1% data, 78.72 F1 10% data	Ontonotes 5.0	ElMo+Glove-> Bi-LSTM	+9.7% F1, +3.4% F1
2019 [35] Logic-RNN	NER Sentiment	Commerce	Type 2	Perf. DE	M Rules POS - D, O - WordNet	KG, POS-> Input	Input relations, mod by masking	Sup, Semi-Sup	Adagrad	Bi-LSTM	81.6 F1, 48.2 Acc 5% data	CoNLL-2003, Yelp 2013	Bi-LSTM, 2020 [57][1], Bi-LSTM	+1.7% F1, -13.7% F1, +4.3% Acc
2019 [36] BO-LSTM	RE	Biomedical	Type 2	Perf.	O - WordNet, GO, HPO, ChEBI, ...	KG, SDP-> Input	MM input relations	Sup	Adam, RMSprop	Word2Vec-> LSTM	75.1 F1	SemEval 2013 task 3 DDI	2018 Hier-LSTM, 2022 [58][1]	+2.2% F1, -9.8% F1
2019 [29] Tree Transformer	MT	Spoken, News	Type 5	Perf.	POS - C	POS-> MHA, Adjecency	MHA mod by POS Tree	Sup	AdamW	Transformer	29.47 BLEU 29.95 BLEU	IWSLT En-De, WMT En-De 2014	2020 Transformer, 2020 DynamicConv, 2021 [47][1]	+3.4% BLEU, +3.2% BLEU, -10.7% BLEU
2020 [41]	Summ	News	Type 2	DE	TF-IDF, O - ConceptNet, KB - Wikipedia	TF-IDF-> Input	Pipeline	Semi-Sup & Prompt tuning	Adam	BART	33.01 R-L <1% data, 36.92 R-L 3.5% data	MA-News (aspect)	BART <1% data, (eq. at 3.5% data)	+21.4% R-L, +0.6% R-L
2020 [49] KERMIT	Sentiment	Reviews	Type 5 (surrogate)	XAI	POS - C, D Relational	POS->Surrogate input	Input Tree-> Surogate	Sup	Adam	XLNet (MHA)	88.99 Acc 88.99 Acc	Yelp polarity Yelp polarity	XLNet 2022 DistilBERT[1]	+7.8% Acc -8.1% Acc
2021 [46] BioKGLM	NER	Biomedical	Type 2	Perf.	OpenKE [59] O - SemMedDB, HGCN, Reactome	KGE-> Input, HL	Representation agg., with KGE	Sup	Adam	BERT	92.34 F1 88.65 F1	BC5CDR BC5CDR-disease	BioBERT 2020 [51][1]	+2.09% F1 +0.17% F1
2021 [28]	NER -> Table	Biomedical	Type 4	Perf.	M Rules, KB - Facts, Cancer Registry	Rules-> Output	Constrained with rules	Sup	n/a	GloVe+RoBERTa-> BiLSTM-CRF	98 F1 NER, 91 F1 Table	n/a	n/a	n/a
2021 [32] BioBERT+MDAtt	QA NLI	Biomedical	Type 2	Perf.	KB - Medical Dependency	KG->MHA Adjecency	MHA mod by masking	Sup	Adam	BioBERT	61 Acc, 84.3 F1	PubMedQA, Evidence Inf. 2.0	2021 BioBERT, 2023 [33][1], 2020 [60][1],	+13.1% F1, -25.2% F1, -0.8% F1
2021 [30]Transformer+LPEA+RHE	MT	News	Type 5	Perf.	POS - C, Rare words	POS->MHA Adjecency	MHA mod to redundand head	Sup	n/a	Transformer (88.1M)	28.27 BLEU	WMT En-De 2014	Transformer, 2021 [47][1]	+3.4% BLEU, -11% BLEU
2022 [48] MetaMap	NER	Biomedical	Type 4	XAI	O - EBRO	KG-> Output	Explaining outputs with O	Sup & HF-Active learning	AdamW	SciBERT	92.87 F1, Human eval.	Medline	2015 MetaMap (symbolic)	+1.0% F1
2022 [31]FRA Transformer	MT	Spoken	Type 5	Perf.	KB - facts	KG->MHA Adjecency	MHA mod by add. facts	Sup	Adam	Transformer	36.1 BLEU	IWSLT14 (De–En)	2019 [29], Tree Transformer	+0.3% BLEU
2022 [43] PTR	RE, NER, Nested NER	General Biomedical	Type 2	Perf.	M Rules	Rules-> Input	Pipeline Markers	Prompt tuning	AdamW	RoBERTa	90.9 F1 90.9 F1, 72.2 F1 supervised	ReTACRED, ReTACRED, OntoNotes 5.0	ENT Marker (Punct) 2021 [61][1], 2020 [44][1]	+0.00% F1 (RE) -0.5% F1 (RE), -22.1% F1 (NER)
2022 [38] SocialBERT	Sentiment	Social texts	Type 2	Perf.	KB - SocialNet Random Walk, SVD	KGE-> Input, HL	Input, mod by agg. layers	Self-Sup	Adam	BERT, RoBERTa	2.62 Perplexity, 1.39 Loss	Author aware LM	BERTbase (20.000 steps)	-18.8% Perplexity[2], -17.7% Loss[2]
2022 [34] Sentic GCN-BERT	Sentiment	General	Type 2	Perf.	POS - D, KB - SenticNet, Lexicon	POS->GNN Adjecency	Adjecency mod with O, D	Sup	Adam	GCN - BERT	85.32 Acc (BERT), 71.28 F1 (BERT)	REST15	DGEDT-BERT	+1.5% Acc, +0.4% F1
2023 [50] GRU-NSR	QA	User Manuals	Type 4	XAI, Perf.	Boing manuals M Rules	Rules-> Output	n/a	Sup	Adam	GRU (RNN)	96.2 Acc, 98.9 BLEU	Boeing manuals	Transformer	+0.2% Acc, +0.3% BLEU,
2023 [45] KEprompt (Dialogue)	Stance detection	Social texts	Type 2	Perf.	M Rule, KB - ConceptGraph, SenticNet,	Rules, KG-> Input	Pipeline	Sup & Prompt tuning	Adam	VAE -> RoBERTa-l	80.47 F1(avg)	SemEval-2016 Task 6	RoBERTa-large-KPT	+3.8% F1
2023 [37] CSK-HGAT	Cls, NLI, Sentiment	Social texts	Type 2	DE	KB - YAGO, Probase, ConceptNet	KG->GNN Adjecency	MM input relations	Semi-Sup	Adam	(ConceptNet Numberbatch, GloVe, BERT) ->GAT	69.15 Acc <1% data, 57.68 Acc <1% data, 70.67 Acc <1% data, 42.87 Acc <1% data	AGNews, SICK, MPQA, TREC	STCKA, BiLSTM-SA, MP-GCN, CNN	+20.7% Acc, +2.4% Acc, +4.5% Acc, +15.37% Acc
2023 [27] PT+BERT	QA, NLI	General	Type 3	DE	PGM O - ConceptNet	Teacher-> Student	Constrained with rules	Semi-Sup	n/a	BERT, GPT2 (MHA)	32.6 Acc 1% data, 34.2 Acc no training	OpenBookQA	BERT, Self-Talk GPT2	+20.8% Acc, +9.9% Acc
2023 [26] Logic-LNCL	NER, Sentiment	News	Type 4	Perf., DE	M Rules	PSL-> Output	Constrained with rules	Sup Noisy labels	EM – Adam, EM - Adadelta	GRU (RNN), CNN	64.06 F1 (50 runs), 79.22 Acc (30 runs)	CoNLL-2003 NER, Polarity	2018 CL(MW,5), 2013 AggNet	+2.9% F1, +0.95% Acc
2023 [39] Toolformer (Dialogue)	QA	General	Type 4-> Type 2	PE	QA, Date, Calculator, Search, MT	LLM Output-> API->2. Input	Pipeline (If API token in top k then use call)	Self-Sup	n/a	GPT-J (6.7B MHA)	33.8 Acc, 17.7 Acc	SQuAD, NQ, NQ	GPT-3 (175B), GPT-3 (175B), Atlas [62][1]	+20.7% Acc, -21.6% Acc, -72.3% Acc
2023 [40] PAL (Dialogue)	QA	General Math	Type 4	PE	Python	LLM Output-> API->2. Input	Pipeline (API call)	Few-shot prompting	n/a	CODEX (15B MHA)	80.4 Acc	GSM8K	CoT (CODEX), 2023 GPT4 [53][1]	+2.9% Acc, -18.3% Acc

[1] year of best result publication with results taken from Papers With Code website.
[2] Lower perplexity and loss is better

Abbreviations: add. - adding, agg. - aggregation of layers, BEAM - Beam search, C - Constituency, Cls - Classification, D - dependency, DE - Data efficiency, EM - Expectation maximization, eval. - evaluation, HF - Human feedback, HL - Hidden layer, I - Interpretability, KI - Knowledge informed, KG - Knowledge graph, KGE - Knowledge graph embedding, LAG - Lagrangian dual form min-max optimization, M Rules - Manually created rules, MM - Multi-modal, mod - modification, O - Ontology, OOD - Out-of-distribution, PE - Parameter efficiency, Perf. - Performance, PGM - Probabilistic graphical models, POS - Part-of-speech tagging, PSL - Probabilistic soft logic, R-L - ROUGE-L, Rare words - is a database with rare occurring words and their meaning, SDP - Shortest dependency path, Sense. - Common sense, SRL - Semantic role labeling, Summ - Summarization task, Sup - Supervised, VAE - Variational auto-encoder and XAI - EXplainable AI.

raises questions of team accountability. Several biomedical use cases, such as a cancer registry pipeline with Bi-LSTM and fact-checking KB [28], and COVID-19 clinical NER with EBRO ontology and BERT LLM [48] were focused on solving real-world problems contributing highly accurate models. A different kind of use case NSAI solution for aeronautic maintenance [50] demanded a creation of dedicated dataset and compared GRU-based RNN and transformer architectures to translate text into KB queries. Producing human-centric AI application for learning and inspecting Boeing aircraft parts and upkeep procedures. Newer methods for LLM, often implemented as an NSAI system, retrieval augmented generation (RAG) [56] reduced hallucinations, although they still carry the risk of the frozen model missuses which demands dedicated attention to bring about trustworthy AI systems.

V. Conclusion

Regularization of neural networks with existing domain expertise from KBs in NLP requires additional data preprocessing and potentially a modification to the deep learning architecture but shows measurable benefits to algorithms. We have shown that NSAI methods exhibit the greatest improvements for data-efficient (DE) AI systems, providing robust solutions when there is little labeled data; these models could be used in assisted annotation tools for various NLP tasks. Additionally, most NSAI methods improved the performance of the DL models for which they were developed but some fell short of their capabilities when compared to the best model.

Efficient fine-tuning of parameters, prompt tuning, and reasoning structures attempt to overcome the cost of training and using LLMs. We have shown that the combinations of symbolic structured information and connectionist models offer a sound alternative that overcomes the critical limitations of individual DL and symbolic models, thus enabling trustworthy and accurate AI systems saving precious limited resources for everyone.

Acknowledgment

This work is supported by the Croatian Science Foundation project: Exploring Interactions Between Regulatory Variants in Human Disease Context (HRZZ−UIP−2020−02 − 1623).

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[2] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[4] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[5] S. Dash, I. Lyngaas, J. Yin, X. Wang, R. Egele, G. Cong, F. Wang, and P. Balaprakash, "Optimizing distributed training on frontier for large language models," *arXiv preprint arXiv:2312.12705*, 2023.

[6] G. Frisoni, G. Moro, and A. Carbonaro, "A survey on event extraction for natural language understanding: Riding the biomedical literature wave," *IEEE Access*, vol. 9, p. 160721 – 160757, 2021, cited by: 23; All Open Access, Gold Open Access, Green Open Access.

[7] A. Oltramari, J. Francis, F. Ilievski, K. Ma, and R. Mirzaee, "Generalizable neuro-symbolic systems for commonsense question answering," *CoRR*, vol. abs/2201.06230, 2022.

[8] J. Yang, G. Xiao, Y. Shen, W. Jiang, X. Hu, Y. Zhang, and J. Peng, "A survey of knowledge enhanced pre-trained models," *CoRR*, vol. abs/2110.00269, 2021.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[10] X. Liu, Z. Lu, and L. Mou, "Weakly supervised reasoning by neuro-symbolic approaches," *Frontiers in Artificial Intelligence and Applications*, vol. 369, p. 665 – 692, 2023, cited by: 0; All Open Access, Green Open Access.

[11] K. Hamilton, A. Nayak, B. Božić, and L. Longo, "Is neuro-symbolic AI meeting its promises in natural language processing? A structured review," *Semantic Web*, no. Preprint, pp. 1–42, 2022.

[12] R. Grishman and B. Sundheim, "Design of the MUC-6 evaluation," in *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.

[13] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3829–3839.

[14] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering," in *Conference on Empirical Methods in Natural Language Processing*, 2018.

[15] C. Elkan and R. Greiner, "Building large knowledge-based systems: Representation and inference in the cyc project: D.b. lenat and r.v. guha," *Artificial Intelligence*, vol. 61, no. 1, pp. 41–52, 1993.

[16] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, p. 39–41, nov 1995.

[17] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: A tool for annotating and analyzing human hereditary disease," *The American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.

[18] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[21] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[22] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[24] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," *arXiv preprint arXiv:2012.09699*, 2020.

[25] Y. Nandwani, A. Pathak, Mausam, and P. Singla, "A primal-dual formulation for deep learning with constraints," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] Z. Chen, H. Sun, H. He, and P. Chen, "Learning from noisy crowd labels with logics," in *Proceedings - International Conference on Data Engineering*, vol. 2023-April, 2023, p. 41 – 52.

[27] S. N. Aakur and S. Sarkar, "Leveraging symbolic knowledge bases for commonsense natural language inference using pattern theory," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, p. 13185 – 13202, 2023.

[28] H.-J. Dai, Y.-H. Yang, T.-H. Wang, Y.-J. Lin, P.-J. Lu, C.-Y. Wu, Y.-C. Chang, Y.-Q. Lee, Y.-C. Zhang, Y.-C. Hsu, H.-H. Wu, C.-R. Ke, C.-J. Huang, Y.-T. Wang, S.-F. Yang, K.-C. Hsiao, K.-J. Liu, L.-T. Chen, I.-S. Chang, K. S. C. Chao, and T.-W. Liu, "Cancer registry coding via hybrid neural symbolic systems in the cross-hospital setting," *IEEE Access*, vol. 9, p. 112081 – 112096, 2021.

[29] X.-P. Nguyen, S. Joty, S. Hoi, and R. Socher, "Tree-structured attention with hierarchical accumulation," in *International Conference on Learning Representations*, 2019.

[30] T. Zhang, H. Huang, C. Feng, and L. Cao, "Enlivening Redundant Heads in Multi-head Self-attention for Machine Translation," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2021, p. 3238 – 3248.

[31] F. Li, J. Zhu, H. Yan, and Z. Zhang, "Grammatically derived factual relation augmented neural machine translation," *Applied Sciences (Switzerland)*, vol. 12, no. 13, 2022.

[32] T. Kang, A. Turfah, J. Kim, A. Perotte, and C. Weng, "A neuro-symbolic method for understanding free-text medical evidence," *Journal of the American Medical Informatics Association*, vol. 28, no. 8, p. 1703 – 1711, 2021.

[33] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami *et al.*, "Meditron-70b: Scaling medical pretraining for large language models," *arXiv preprint arXiv:2311.16079*, 2023.

[34] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," *Knowledge-Based Systems*, vol. 235, p. 107643, 2022.

[35] B. Chen, Z. Hao, X. Cai, R. Cai, W. Wen, J. Zhu, and G. Xie, "Embedding logic rules into recurrent neural networks," *IEEE Access*, vol. 7, pp. 14 938–14 946, 2019.

[36] A. Lamurias, D. Sousa, L. A. Clarke, and F. M. Couto, "BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019.

[37] M. Wu, "Commonsense knowledge powered heterogeneous graph attention networks for semi-supervised short text classification," *Expert Systems with Applications*, vol. 232, 2023.

[38] I. Karpov and N. Kartashev, "SocialBERT – Transformers for Online Social Network Language Modelling," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13217 LNCS, p. 56 – 70, 2022.

[39] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *arXiv preprint arXiv:2302.04761*, 2023.

[40] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, "Pal: Program-aided language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 764–10 799.

[41] B. Tan, L. Qin, E. Xing, and Z. Hu, "Summarizing text on any aspects: A knowledge-informed weakly-supervised approach," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 6301–6309.

[42] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[43] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "PTR: Prompt Tuning with Rules for Text Classification," *AI Open*, vol. 3, p. 182 – 192, 2022.

[44] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 465–476.

[45] H. Huang, B. Zhang, Y. Li, B. Zhang, Y. Sun, C. Luo, and C. Peng, "Knowledge-enhanced prompt-tuning for stance detection," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 6, 2023, cited by: 1.

[46] H. Fei, Y. Ren, Y. Zhang, D. Ji, and X. Liang, "Enriching contextualized language model from knowledge graph for biomedical information extraction," *Briefings in Bioinformatics*, vol. 22, no. 3, 2021, cited by: 42.

[47] S. Takase and S. Kiyono, "Lessons on parameter sharing across layers in transformers," *arXiv preprint arXiv:2104.06022*, 2021.

[48] M. Arguello-Casteleiro, C. Henson, N. Maroto, S. Li, J. Des-Diz, M. Fernandez-Prieto, S. Peters, T. Furmston, C. Sevillano Torrado, D. Maseda Fernandez, M. Kulshrestha, J. Keane, R. Stevens, and C. Wroe, "MetaMap versus BERT models with explainable active learning: Ontology-based experiments with prior knowledge for COVID-19," vol. 3127, 2022, Conference paper, p. 108 – 117, cited by: 2.

[49] F. M. Zanzotto, A. Santilli, L. Ranaldi, D. Onorati, P. Tommasino, and F. Fallucchi, "KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020, p. 256 – 267.

[50] A. Siyaev, D. Valiev, and G.-S. Jo, "Interaction with Industrial Digital Twin Using Neuro-Symbolic Reasoning," *Sensors*, vol. 23, no. 3, 2023.

[51] H.-C. Shin, Y. Zhang, E. Bakhturina, R. Puri, M. Patwary, M. Shoeybi, and R. Mani, "BioMegatron: Larger biomedical domain language model," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4700–4706.

[52] M. Besta, F. Memedi, Z. Zhang, R. Gerstenberger, N. Blach, P. Nyczyk, M. Copik, G. Kwaśniewski, J. Müller, L. Gianinazzi, A. Kubicek, H. Niewiadomski, O. Mutlu, and T. Hoefler, "Topologies of reasoning: Demystifying chains, trees, and graphs of thoughts," 2024.

[53] A. Zhou, K. Wang, Z. Lu, W. Shi, S. Luo, Z. Qin, S. Lu, A. Jia, L. Song, M. Zhan *et al.*, "Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification," *arXiv preprint arXiv:2308.07921*, 2023.

[54] D. Horvatić and T. Lipic, "Human-Centric AI: The Symbiosis of Human and Artificial Intelligence," *Entropy*, vol. 23, no. 3, 2021.

[55] A. Krajna, M. Brcic, T. Lipic, and J. Doncevic, "Explainability in reinforcement learning: perspective and position," *arXiv preprint arXiv:2203.11547*, 2022.

[56] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[57] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu, "Automated concatenation of embeddings for structured prediction," *arXiv preprint arXiv:2010.05006*, 2020.

[58] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining language models with document links," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8003–8016.

[59] X. Han, S. Cao, L. Xin, Y. Lin, Z. Liu, M. Sun, and J. Li, "OpenKE: An Open Toolkit for Knowledge Embedding," in *Proceedings of EMNLP*, 2018.

[60] J. DeYoung, E. Lehman, B. Nye, I. J. Marshall, and B. C. Wallace, "Evidence inference 2.0: More data, better models," *arXiv preprint arXiv:2005.04177*, 2020.

[61] S. Park and H. Kim, "Improving sentence-level relation extraction through curriculum learning," *arXiv preprint arXiv:2107.09332*, 2021.

[62] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Few-shot learning with retrieval augmented language models," *arXiv preprint arXiv:2208.03299*, 2022.