Investigating Logic Tensor Networks for Neural-Symbolic Argument Mining

调查逻辑张量网络在神经符号论证挖掘中的应用

Andrea Galassi, [1] Marco Lippi, [2] Paolo Torroni [1]

Andrea Galassi, [1] Marco Lippi, [2] Paolo Torroni [1]

[1] Department of Computer Science and Engineering (DISI), University of Bologna, Bologna, Italy

[1] 博洛尼亚大学计算机科学与工程系 (DISI)，意大利博洛尼亚

[2] Department of Sciences and Methods for Engineering (DISMI), University of Modena and Reggio Emilia, Modena, Italy

[2] 摩德纳与雷焦艾米利亚大学工程科学与方法系 (DISMI)，意大利摩德纳

a.galassi@unibo.it, marco.lippi@unimore.it, paolo.torroni@unibo.it

a.galassi@unibo.it, marco.lippi@unimore.it, paolo.torroni@unibo.it

Abstract

摘要

We present an application of neural-symbolic learning to the task of argument mining, where argument components and their relations are extracted from unstructured textual corpora. We use the framework of Logic Tensor Networks to train neural models to jointly fit the data and satisfy specific domain rules. Our experiments on a corpus of scientific abstracts indicate that including symbolic rules during the training process improves classification performance, compliance to the rules, and robustness of the results. As in the case of other neural-symbolic applications, we further discuss how scalability remains a crucial issue.

我们展示了神经符号学习在论证挖掘任务中的应用，其中论证组件及其关系从非结构化文本语料库中提取。我们使用逻辑张量网络的框架来训练神经模型，以共同拟合数据并满足特定领域规则。我们在一组科学摘要的语料库上的实验表明，在训练过程中包含符号规则可以提高分类性能、规则遵从性和结果的稳健性。与其他神经符号应用的情况一样，我们进一步讨论了可扩展性仍然是一个关键问题。

Introduction

引言

Argument Mining (AM) stemmed from Natural Language Processing (NLP) and Knowledge Representation and Reasoning (Cabrio and Villata 2018), with the goal of automatically extracting arguments and their relations from natural language texts (Lippi and Torroni 2016).

论证挖掘 (AM) 源于自然语言处理 (NLP) 和知识表示与推理 (Cabrio 和 Villata 2018)，其目标是自动从自然语言文本中提取论证及其关系 (Lippi 和 Torroni 2016)。

Argumentation is an ancient discipline that has its roots in logic and philosophy, aimed to study the way in which humans debate and reason. Inspired by the seminal work of Dung (1995), the application of computer science to the domain of argumentation has brought to the development of a fertile research area named computational argumentation. While several definitions of argument exist in the literature, one of the most intuitive has been given by Douglas Walton (2009): an argument is defined as a statement about a topic, usually named claim, possibly supported by a set of premises. The discipline of AM aims to extract such argument components from textual corpora, as well as the relations between them, which can be, for example, a supporting relation between a premise and a claim, or an attacking relation between two different claims.

论证是一门古老的学科，起源于逻辑和哲学，旨在研究人类辩论和推理的方式。受邓 (Dung，1995) 开创性工作的启发，将计算机科学应用于论证领域催生了一个名为计算论证的肥沃研究领域。尽管文献中存在多种论证的定义，但道格拉斯 · 沃尔顿 (Douglas Walton，2009) 给出的定义是最直观的: 论证被定义为关于某一主题的陈述，通常称为主张，可能由一组前提支持。AM(Argument Mining) 学科旨在从文本语料库中提取这些论证组件及其之间的关系，例如，前提与主张之间的支持关系，或两个不同主张之间的攻击关系。

Like in most NLP applications, deep learning has recently pushed the state-of-the-art also in AM. Yet, many challenges still stand open, as argumentation involves tasks such as reasoning, debate and persuasion that cannot be easily addressed by deep architectures only, sophisticated as they may be. For that reason, Galassi et al. (2019) argue that a combination of symbolic and sub-symbolic approaches could leverage significant advances in AM by exploiting domain knowledge and the constraints imposed by the underlying argument model. They illustrate that idea using two neural-symbolic (NeSy) frameworks, DEEP-PROBLOG (Manhaeve et al. 2021) and Grounding-Specific Markov Logic Networks (Lippi and Frasconi 2009), but do not offer empirical evaluations. Unfortunately, many of the existing NeSy frameworks are under continuous development and their applications are often limited to a single domain and a few case studies. In particular, NLP tasks are seldom considered, though we believe they represent important and challenging benchmarks.

与大多数自然语言处理 (NLP) 应用一样，深度学习最近也推动了 AM 的最新进展。然而，许多挑战仍然存在，因为论证涉及推理、辩论和说服等任务，这些任务仅靠深度架构难以轻易解决，尽管它们可能非常复杂。因此，Galassi 等 (2019) 认为，符号与亚符号方法的结合可以通过利用领域知识和基础论证模型施加的约束，在 AM 中带来显著进展。他们使用两个神经符号 (NeSy) 框架，DEEP-PROBLOG(Manhaeve 等，2021) 和特定基础的马尔可夫逻辑网络 (Lippi 和 Frasconi，2009) 来说明这一观点，但未提供实证评估。不幸的是，许多现有的 NeSy 框架仍在持续开发中，其应用通常仅限于单一领域和少数案例研究。特别是，NLP 任务很少被考虑，尽管我们认为它们代表了重要且具有挑战性的基准。

In this vein, Pacheco and Goldwasser (2021) analyze existing NeSy frameworks, observing that they are not specifically designed to support a variety of NLP tasks, and critically lack of a series of important features. We shall add to the list of shortcomings a lack of support for collective classification (Sen et al. 2008). This is a fundamental feature for AM, since argument analysis is exquisitely context-dependent, and the task of classifying a single argumentative component (or relation) should be carried out by considering not only the attributes of that component or relation, but also of the attributes of other connected components and relations. To address these limitations, Pacheco and Goldwasser (2021) introduce the DRAIL NeSy framework and show its application in the AM domain. To the best of our knowledge, no other NeSy approach to AM has been investigated so far.

在这方面，Pacheco 和 Goldwasser (2021) 分析了现有的 NeSy 框架，观察到它们并不是专门设计来支持各种 NLP 任务，并且严重缺乏一系列重要特性。我们将缺乏对集体分类的支持 (Sen et al. 2008) 加入到缺点列表中。这是 AM 的一个基本特征，因为论证分析极其依赖于上下文，分类单个论证组成部分 (或关系) 的任务不仅应考虑该组成部分或关系的属性，还应考虑其他连接组成部分和关系的属性。为了解决这些局限性，Pacheco 和 Goldwasser (2021) 引入了 DRAIL NeSy 框架，并展示了其在 AM 领域的应用。根据我们所知，目前尚未有其他 NeSy 方法对 AM 进行研究。

In this work, we address AM using a different NeSy framework, namely Logic Tensor Networks (Serafini and d'Avila Garcez 2016). We focus on the classification of argumentative component and prediction of links between component pairs. Importantly, LTNs allow us to easily decouple the symbolic and sub-symbolic parts of the model, and enable collective classification during training. Our results indicate that the introduction of logic rules improves classification performance, compliance to the rules, and robustness of the results. To the best of our knowledge, this is also the first application of LTNs to NLP.

在这项工作中，我们使用不同的 NeSy 框架来解决 AM，即逻辑张量网络 (Serafini 和 d'Avila Garcez 2016)。我们专注于论证组成部分的分类和组件对之间链接的预测。重要的是，LTNs 使我们能够轻松解耦模型的符号部分和子符号部分，并在训练过程中实现集体分类。我们的结果表明，引入逻辑规则改善了分类性能、规则遵从性和结果的稳健性。根据我们所知，这也是 LTNs 在 NLP 中的首次应用。

Logic Tensor Networks (LTNs)

逻辑张量网络 (LTNs)

Logic Tensor Networks (LTNs) (Serafini and d'Avila Garcez 2016; Donadello, Serafini, and Garcez 2017; Badreddine et al. 2020) are a framework that integrates first-order many-valued logical reasoning (Bergmann 2008) with tensor networks (Socher et al. 2013), implemented in Tensor-Flow (Abadi et al. 2016). LTNs belong to the "tensoriza-tion" class of undirect NeSy approaches (De Raedt et al. 2020) which embed First-Order Logic (FOL) entities, such as constants and facts, into real-valued tensors. The framework enables to combine data-driven machine learning with background knowledge expressed through first-order fuzzy logic representations. Therefore, one can use FOL to impose soft constraints at training time and investigate properties at test time. Once trained, neural architectures can be used independently from the framework.

逻辑张量网络 (LTNs) (Serafini 和 d'Avila Garcez 2016; Donadello, Serafini, 和 Garcez 2017; Badreddine 等 2020) 是一个将一阶多值逻辑推理 (Bergmann 2008) 与张量网络 (Socher 等 2013) 结合的框架，使用 TensorFlow (Abadi 等 2016) 实现。LTNs 属于无向神经系统方法的“张量化”类别 (De Raedt 等 2020)，将一阶逻辑 (FOL) 实体，如常量和事实，嵌入到实值张量中。该框架使得可以将数据驱动的机器学习与通过一阶模糊逻辑表示的背景知识相结合。因此，可以在训练时使用 FOL 强加软约束，并在测试时研究属性。一旦训练完成，神经架构可以独立于该框架使用。

LTN variables are an abstract representation of data. They must be linked to a set of real-valued vectors, which are all the possible groundings of that variable. A single data point of this set can be represented using LTN constants. LTN functions represent operations over variables and produce real-valued vectors. The evaluation is done by a set of Ten-sorFlow operations, e.g., a neural network, defined together with the function. LTN predicates are a special class of functions whose output is a single real value between 0 and 1, which represents the degree of truth of the predicate. They can be used to represent classes of objects as well as properties that may hold between multiple objects. The learning setting is defined in terms of LTN axioms, i.e., formulas that specify logic conditions in terms of predicates, functions, and variables and can be used to assign labels to data and to specify soft constraints. Axioms can include logical connectives $(\wedge, \vee, \sim, \Rightarrow)$[1] and quantifiers $(\forall, \exists)$.

LTN 变量是数据的抽象表示。它们必须与一组实值向量相联系，这些向量是该变量的所有可能基础。该集合的单个数据点可以使用 LTN 常量表示。LTN 函数表示对变量的操作，并生成实值向量。评估是通过一组 TensorFlow 操作完成的，例如，与函数一起定义的神经网络。LTN 谓词是一类特殊的函数，其输出是介于 0 和 1 之间的单个实值，表示谓词的真实性程度。它们可以用于表示对象的类别以及可能在多个对象之间成立的属性。学习设置是通过 LTN 公理定义的，即以谓词、函数和变量为基础的逻辑条件公式，可用于为数据分配标签并指定软约束。公理可以包括逻辑连接词 $(\wedge, \vee, \sim, \Rightarrow)$[1] 和量词 $(\forall, \exists)$ 。

LTNs, similarly to DEEPPROBLOG, enable the creation of vertical-hybrid learning systems, where high-level logic is placed on top of deep networks, as opposed to horizontal-hybrid learning (e.g., the work of Hu et al. (2016)), where the symbolic knowledge is encoded into the networks themselves (d'Avila Garcez et al. 2019). The idea behind the design of these systems is that the symbolic part must influence the behavior of the neural part and provide means to interpret their results.

LTN，类似于 DEEPPROBLOG，能够创建垂直混合学习系统，其中高级逻辑位于深度网络之上，而不是水平混合学习 (例如，Hu 等人 (2016) 的工作)，在这种学习中，符号知识被编码到网络本身中 (d'Avila Garcez 等人 2019)。这些系统设计背后的理念是，符号部分必须影响神经部分的行为，并提供解释其结果的手段。

Reasoning is performed in the form of approximate satisfiability, which means that the optimization process aims to maximize the level of satisfiability of a grounded theory, by minimizing the loss function (Serafini and d'Avila Garcez 2016). Inference follows a model-theoretic perspective, which means that learning is done via shared parameters over the ground model, whereas inference is based on possible groundings of the model (De Raedt et al. 2020). After training, it is possible to evaluate queries expressed in FOL, as a means to assess the performance of the neural networks as well as to verify the degree of truth of a property.

推理以近似可满足性的形式进行，这意味着优化过程旨在通过最小化损失函数来最大化一个基础理论的可满足性水平 (Serafini 和 d'Avila Garcez 2016)。推理遵循模型理论的视角，这意味着学习是通过基础模型上的共享参数进行的，而推理则基于模型的可能基础 (De Raedt 等 2020)。训练后，可以评估以一阶逻辑 (FOL) 表达的查询，以评估神经网络的性能以及验证属性的真实性程度。

Argument Mining with LTNs

使用逻辑张量网络进行论证挖掘

We frame both component classification and link prediction as classification tasks. To address them, we define two neural networks, NNCOMP and NNLINK. The first network takes a component in input and produces a probability distribution over the possible component classes. The second receives two components and outputs a single value between 0 and 1, which represents the probability of there being an argumentative link between them. [2]

我们将组件分类和链接预测框定为分类任务。为了解决这些问题，我们定义了两个神经网络，NNCOMP 和 NNLINK。第一个网络以一个组件作为输入，并生成一个关于可能组件类别的概率分

布。第二个网络接收两个组件，并输出一个介于 0 和 1 之间的单一值，表示它们之间存在论证链接的概率。[2]

Data-driven optimization is defined through three elements for each class of both tasks: a variable, a predicate, and an axiom. The predicate is linked to the respective output of our networks, whereas the variable is associated to all the data of the training set that belong to that class, and the axiom combines the previous elements and defines the optimization objective. For example, given a class "CLAIM" of components, we define the variable $vClaim$, the predicate $CLAIM$ and the following axiom:

数据驱动的优化通过每个任务类别的三个元素定义: 一个变量、一个谓词和一个公理。谓词与我们网络的相应输出相关联，而变量与属于该类别的训练集的所有数据相关联，公理结合了前面的元素并定义了优化目标。例如，给定一个“主张”类别的组件，我们定义变量 $vClaim$ 、谓词 $CLAIM$ 和以下公理:

$$\forall v \text{ Claim} : \text{CLAIM } (v \text{ Claim }) \tag{1}$$

The rule-driven optimization is defined through variables linked to all the training data and through specific axioms that express the rules. For example, to enforce the antisymmetric property of links we define two variables (vC_1 and vC_2) and associate them to all the components of the training set, and specify the following axiom:

基于规则的优化通过与所有训练数据相关联的变量和表达规则的特定公理来定义。例如，为了强制链接的反对称属性，我们定义两个变量 (vC_1 和 vC_2 ，并将它们与训练集的所有组件关联，并指定以下公理:

$$\forall vC_1, vC_2 :$$

$$\text{LINK}\,(vC_1, vC_2) \Rightarrow \sim \text{LINK}\,(vC_2, vC_1) \tag{2}$$

Experimental Setting

实验设置

The implementation of LTNs we used does not expose APIs to easily configure some aspects of the training procedure. [3] Indeed, to guarantee the consistency of the tensor network, the training procedure employed in our experiments does not use mini-batches, which unfortunately has repercussions on the computational resources required. This limit is neither theoretical nor methodological, but it derives from the current implementation of the framework: when a predicate is defined in an LTN over a set of variables, all the possible groundings of such variables are used as part of the same batch. This is necessary for the LTN to evaluate the degree of truth of the predicate. Unfortunately, there is no way to easily construct a distinct mini-batch for each different document, therefore even a few simple rules that connect multiple entities, such as the rule shown in Equation 2, being necessarily applied to any component pair in the corpus, are sufficient to make all the data belong to the same batch.

我们使用的 LTN 实现并未暴露 API，以便轻松配置训练过程的某些方面。[3] 实际上，为了保证张量网络的一致性，我们实验中采用的训练过程不使用小批量，这不幸地对所需的计算资源产生了影响。这个限制既不是理论上的，也不是方法论上的，而是源于框架的当前实现: 当在一组变量上定义一个谓词时，所有可能的变量实例都作为同一批次的一部分被使用。这对于 LTN 评估谓词的真值程度是必要的。不幸的是，没有简单的方法为每个不同的文档构建一个独特的小批量，因此，即使是一些简单的规则，将多个实体连接起来，例如方程 2 中所示的规则，必然应用于语料库中的任何组件对，也足以使所有数据属于同一批次。

Due to this scalability issue, we have chosen to experiment on the AbstRCT corpus, which has a limited number of documents, to represent sentences using sentence embed-dings of small size, and to use neural architectures with a reduced number of trainable parameters.

由于这个可扩展性问题，我们选择在 AbstRCT 语料库上进行实验，该语料库文档数量有限，使用小尺寸的句子嵌入表示句子，并使用具有较少可训练参数的神经架构。

Data

数据

The AbstRCT Corpus (Mayer, Cabrio, and Villata 2020; Mayer et al. 2021) consists of 659 abstracts of scientific papers regarding randomized control trials for the treatment of specific diseases. [4] The corpus is divided into three topical

AbstRCT 语料库 (Mayer, Cabrio 和 Villata 2020；Mayer 等 2021) 由 659 篇关于特定疾病治疗的随机对照试验的科学论文摘要组成。[4] 该语料库分为三个主题

Dataset Split	Neoplasm			Glaucoma Test	Mixed Test
	Train	Valid.	Test		
Documents	350	50	100	100	100
Components	2,267	326	686	594	600
Evidence	1,537	218	438	404	338
Claim	730	108	248	190	212
Couples	14,286	2,030	4,380	3,332	3,332
Links	1,418	219	424	367	329

数据集划分	新生物			青光眼测试	混合测试
	训练	验证。	测试		
文档	350	50	100	100	100
组件	2,267	326	686	594	600
证据	1,537	218	438	404	338
主张	730	108	248	190	212
夫妻	14,286	2,030	4,380	3,332	3,332
连接	1,418	219	424	367	329

Table 1: AbstRCT dataset composition.

表 1:AbstRCT 数据集组成。

datasets: neoplasm, glaucoma, and mixed. Neoplasm contains 500 abstracts divided into training (350), test (100), and validation (50) splits. The other two datasets contain 100 abstract each and are designed to be test sets. [5] To the best of our knowledge, this is the only corpus for AM that offers three test sets, allowing general evaluation.

数据集: 肿瘤、青光眼和混合。肿瘤数据集包含 500 篇摘要，分为训练集 (350)、测试集 (100) 和验证集 (50)。另外两个数据集各包含 100 篇摘要，旨在作为测试集。[5] 据我们所知，这是唯一一个为 AM 提供三个测试集的语料库，允许进行全面评估。

The corpus contains about 4,000 argumentative components divided into two classes: EVIDENCE (2,808) and CLAIM (1,390). Out of the almost 25,000 possible pairs of components that belong to the same document, about 10% are connected through a direct link. Its composition is reported in Table 1. The argumentative model chosen for annotation enforces only one constraint: claims can have an outgoing link only to other claims.

该语料库包含约 4,000 个论证组件，分为两个类别: 证据 (2,808) 和主张 (1,390)。在几乎 25,000 个可能属于同一文档的组件对中，约有 10% 通过直接链接相连。其组成在表 1 中报告。所选择的论证模型在注释中仅施加一个约束: 主张只能有指向其他主张的外部链接。

Sentence embeddings are created using pre-trained GloVe embeddings of size 25 (Pennington, Socher, and Manning 2014), by averaging over the words of the sentence. [6] Such a simple method yields a low-dimensional representation without requiring to train new embeddings or rely on dimensionality reduction techniques. In the future we want to investigate more advanced sentence embeddings such as those presented by Reimers and Gurevych (2019) and Cer et al.

[1] The symbol $\sim$ indicates logical negation.

[1] 符号 $\sim$ 表示逻辑否定。

[2] This definition assumes that there is a single type of link. Otherwise, one should simply augment the number of output neurons of nnLink in order to match the number of possible relations.

[2] 该定义假设只有一种类型的链接。否则，应简单地增加 nnLink 的输出神经元数量，以匹配可能关系的数量。

[3] We used version 1.0 of the framework.

[3] 我们使用了该框架的 1.0 版本。

[4] The corpus is available at https://gitlab.com/tomaye/abstrct.

[4] 该语料库可在 https://gitlab.com/tomaye/abstrct 获取。

句子嵌入是使用预训练的 GloVe 嵌入 (大小为 25)(Pennington, Socher, 和 Manning 2014) 创建的，通过对句子中的单词进行平均。[6] 这样一种简单的方法产生了低维表示，而无需训练新的嵌入或依赖于降维技术。未来我们希望研究更高级的句子嵌入，例如 Reimers 和 Gurevych(2019) 以及 Cer 等人提出的嵌入。(2018).

Architecture

架构

For what concerns the neural architecture, we rely on a simple network. The aforementioned scalability issues have prevented us to experiment with NLP state-of-the-art models such as the Transformer (Vaswani et al. 2017) or BERT-based models (Devlin et al. 2019).

关于神经架构，我们依赖于一个简单的网络。上述可扩展性问题阻止我们尝试 NLP 最先进的模型，如 Transformer(Vaswani 等人 2017) 或基于 BERT 的模型 (Devlin 等人 2019)。

Our architecture is made of three stacked fully-connected layers of size 10,20, and 10, followed by a softmax classification layer. We use ReLU as activation function, and employ dropout with probability $p = 0.4$ after each layer. The two models have 712 (NNCOMP) and 962 (NNLINK) trainable parameters. To obtain more robust results with respect to the non-deterministic elements of the training procedure (Goodfellow, Bengio, and Courville 2016), we follow Galassi, Lippi, and Torroni (2021) and train an ensemble of 20 networks both for NNCOMP and NNLINK, and evaluate the aggregated output. Majority voting (MAJ) could be a simple aggregation method. However, that would provide a categorical output, losing the probabilistic semantic of the prediction. That could be a drawback. An alternative would be to use the average of the output of the networks (AVG). That, however, would be vulnerable to outliers. We have therefore decided to try both approaches, so as to better evaluate the options from multiple perspectives. The MAJ and AVG approaches are two among the commonest aggregation methods: there are of course others. In particular, a possibility we plan to explore in future work is to represent the probability score assigned to a class as the percentage of networks that give it the highest probability. Such a method should guarantee robustness while fitting the fuzzy logic semantic of the framework.

我们的架构由三层堆叠的全连接层组成，大小分别为 10、20 和 10，后面跟着一个 softmax 分类层。我们使用 ReLU 作为激活函数，并在每层之后采用概率为 $p = 0.4$ 的 dropout。两个模型分别有 712 个 (NNCOMP) 和 962 个 (NNLINK) 可训练参数。为了获得更稳健的结果，以应对训练过程中的非确定性元素 (Goodfellow, Bengio, and Courville 2016)，我们遵循 Galassi, Lippi, 和 Torroni(2021) 的做法，为 NNCOMP 和 NNLINK 训练 20 个网络的集成，并评估聚合输出。多数投票 (MAJ) 可能是一种简单的聚合方法。然而，这将提供一个分类输出，失去预测的概率语义，这可能是一个缺点。另一种选择是使用网络输出的平均值 (AVG)。然而，这种方法容易受到异常值的影响。因此，我们决定尝试这两种方法，以便从多个角度更好地评估选项。MAJ 和 AVG 方法是最常见的聚合方法之一: 当然还有其他方法。特别是，我们计划在未来的工作中探索的一种可能性是将分配给一个类别的概率分数表示为给予其最高概率的网络的百分比。这种方法应能保证稳健性，同时符合框架的模糊逻辑语义。

Method

方法

To properly evaluate whether the use of symbolic rules within the model yields positive results, we compare against a baseline model where only the sub-symbolic component is exploited. In the NeSy model, we include two LTN axioms based on characteristic properties of the corpus: (i) no symmetric link can exist, and (ii) claims can be linked only to other claims. For the purely data-driven approach, we make use of three predicates, corresponding to the classes of the dataset: LINK, EVIDENCE, and CLAIM. For our NeSy approach we include the axioms reported in Equations 2 and 3. In particular, the latter axiom connects the two tasks, thus inducing a joint-learning setting.

为了正确评估模型中使用符号规则是否产生积极结果，我们与仅利用子符号组件的基线模型进行比较。在 NeSy 模型中，我们包括了基于语料库特征属性的两个 LTN 公理:(i) 不存在对称链接，以及 (ii) 断言只能与其他断言链接。对于纯数据驱动的方法，我们使用三个谓词，分别对应数据集的类别:LINK、EVIDENCE 和 CLAIM。对于我们的 NeSy 方法，我们包括了方程 2 和 3 中报告的公理。特别是，后一个公理连接了这两个任务，从而引入了联合学习的设置。

$$\forall vC_1, vC_2 : \mathrm{LINK}\,(vC_1, vC_2)$$

$$\wedge \mathrm{CLAIM}\,(vC_1) \Rightarrow \mathrm{CLAIM}\,(vC_2) \tag{3}$$

To avoid overfitting, we early-stop the process by monitoring the F1 score of link prediction on the validation set, using patience of 1,000 epochs. We intentionally focus on link prediction because it is considered the most challenging task, and arguably the one that would benefit the most from the introduction of rules.

为了避免过拟合，我们通过监控验证集上链接预测的 F1 分数来提前停止该过程，耐心值设为 1,000 个周期。我们故意专注于链接预测，因为它被认为是最具挑战性的任务，并且可以说是最能从规则引入中受益的任务。

We evaluate the two models (neural baseline and NeSy) along the following dimensions:

我们从以下几个维度评估这两个模型 (神经基线和 NeSy):

- Performance: we measure the F1 score for the tasks of link prediction and component classification, to assess whether the injection of rules contributes to improve the performance of the models;

- 性能: 我们测量链接预测和组件分类任务的 F1 分数，以评估规则的注入是否有助于提高模型的性能；

- Compliance: we test whether the models respect the two desired properties, through LTN queries performed on the AVG ensemble;

- 合规性: 我们通过在 AVG 集合上执行 LTN 查询来测试模型是否遵循这两个期望的属性；

- Robustness: we compute the degree of agreement between the networks related to the predictions of the MAJ ensemble, to assess whether the use of rules increase robustness against the intrinsic randomness of the training process.

- 鲁棒性: 我们计算与 MAJ 集合的预测相关的网络之间的一致性程度，以评估规则的使用是否提高了对训练过程内在随机性的鲁棒性。

Our experimental evaluation does not include a comparison with other neuro-symbolic methods, since no software can be easily used as an off-the-shelf competitor representing neuro-symbolic state-of-the-art approaches.

我们的实验评估不包括与其他神经符号方法的比较，因为没有软件可以轻松地作为现成的竞争者，代表神经符号的最先进方法。

As for NLP state-of-the-art models, we decided to not include results obtained by previous neural approaches in our tables, considering them misleading for the purpose of this work. Indeed, our aim is to show the positive effect of the use of a neuro-symbolic framework over a plain neural architecture for a very challenging NLP task such as AM, by introducing symbolic rules as a driving element of the training procedure.

至于自然语言处理的最先进模型，我们决定不在我们的表格中包含先前神经方法获得的结果，因为我们认为这些结果对本工作的目的具有误导性。实际上，我们的目标是通过引入符号规则作为训练过程的驱动元素，展示使用神经符号框架相对于普通神经架构在诸如自动标注等非常具有挑战性的自然语言处理任务中的积极效果。

State-of-the-art models would have an inherent advantage against ours due to their large size (millions of parameters against less than one thousand), and therefore the comparison would not provide any useful information. In the future, after addressing the scalability issues, we aim to include such architectures in our experiments, by comparing their standard training mechanism against their training using LTNs.

由于其庞大的规模 (数百万个参数，而我们的模型不到一千个)，最先进的模型相对于我们的模型具有固有的优势，因此比较不会提供任何有用的信息。在未来，在解决可扩展性问题后，我们计划在实验中包含此类架构，通过比较其标准训练机制与使用逻辑张量网络 (LTNs) 的训练。

---

[5] Some of the documents of the neoplasm and glaucoma test set are also included into the mixed set.

[5] 一些新生物和青光眼测试集的文档也被包含在混合集中。

[6] GloVe word embeddings can be downloaded at https://nlp.stanford.edu/projects/glove/.

[6] GloVe 词嵌入可以在 https://nlp.stanford.edu/projects/glove/ 下载。

Infrastructure and Runtime Details

基础设施和运行时细节

We have performed all our experiments on the following infrastructure: ASRock Z370 Pro4 motherboard, GeForce GTX 1080 Ti GPU, Intel Core i7-8700K @ 3.70GHz CPU. Using the baseline approach, the average training time for each network is less than one minute. Using our NeSy approach, the average training time for each network is 14 minutes, with a standard deviation of about 3 minutes. Inference can be performed on the whole ensemble of 20 networks in less than 30 seconds in all the considered test datasets and for all the considered approaches.

我们在以下基础设施上进行了所有实验:ASRock Z370 Pro4 主板，GeForce GTX 1080 Ti GPU，Intel Core i7-8700K @ 3.70GHz CPU。使用基线方法，每个网络的平均训练时间不到一分钟。使用我们的神经符号方法，每个网络的平均训练时间为 14 分钟，标准偏差约为 3 分钟。在所有考虑的测试数据集和所有考虑的方法中，推理可以在少于 30 秒内对 20 个网络的整个集成进行。

Results

结果

Table 2 summarizes the results of our experiments. For the classification tasks, we report the macro-F1 score for component classification and the F1 score for the link class. The agreement is measured as Krippendorff's α , while the degree of truth of the properties is evaluated through LTN queries. For what concerns the AM tasks, the difference between the MAJ and AVG approaches is negligible in the rule-based setting, while it is more evident in the no-rules setting for link prediction, where the majority voting achieves better performance.

表 2 总结了我们实验的结果。对于分类任务，我们报告了组件分类的宏 F1 分数和链接类别的 F1 分数。协议通过 Krippendorff's α 测量，而属性的真实性程度通过 LTN 查询进行评估。关于 AM 任务，MAJ 和 AVG 方法之间的差异在基于规则的设置中微不足道，而在无规则设置下的链接预测中则更为明显，其中多数投票取得了更好的性能。

The use of rules seems to be beneficial especially for the task of link prediction, where the networks perform consistently better than those trained without rules. Conversely, in a few cases the latter perform marginally better on component classification. The results are, however, comparable.

规则的使用似乎对链接预测任务特别有利，网络的表现明显优于那些没有规则训练的网络。相反，在少数情况下，后者在组件分类上表现略好。然而，结果是可比的。

The use of rules clearly benefits robustness, boosting the agreement by at least 5 points for link prediction and a few points for component classification. This benefit is confirmed also by the smaller difference between the AVG and the MAJ approaches for classification.

规则的使用显然有助于增强鲁棒性，使链接预测的协议提高至少 5 个百分点，组件分类则提高了几个百分点。这一好处也通过分类中 AVG 和 MAJ 方法之间较小的差异得到了证实。

Finally, as far as compliance to the rules, we observe that a purely data-driven approach already satisfies the properties almost completely. However, the introduction of rules during training further improves compliance, pushing it very close to 100% . All these results appear to be consistent across the three test sets.

最后，就规则遵从性而言，我们观察到纯数据驱动的方法几乎完全满足属性。然而，在训练过程中引入规则进一步改善了遵从性，使其非常接近 100% 。所有这些结果在三个测试集上似乎是一致的。

The high value of compliance may seem unusual, but it can be easily explained. A logic clause $A \Rightarrow B$ is considered true when both A and B are true or when A is false. The high value obtained by the base method are partially due to the latter case. Indeed, Eqs. 2 and 3 have a LINK predicate in their left part, so for every pair of components the equation will result true for every case where they are not linked. Since in each test set the number of linked pairs amount to about 10% of the total, the lower bound in respecting the rules is around 90% . Finally, we remark that an improvement of 1-2 percentage points corresponds to 30-40 pairs of components, which we believe to be not negligible.

合规性的高值可能看起来不寻常，但这可以很容易地解释。逻辑子句 $A \Rightarrow B$ 在 A 和 B 都为真或 A 为假时被认为是真。基础方法获得的高值部分是由于后者的情况。实际上，方程 2 和 3 的左侧有一个 LINK 谓词，因此对于每一对组件，当它们没有链接时，方程将为真。由于在每个测试集中，链接对的数量约占总数的 10% ，因此遵守规则的下限约为 90% 。最后，我们指出，1-2 个百分点的改进对应于 30-40 对组件，我们认为这并非微不足道。

Discussion

讨论

We presented the first application of LTNs to a challenging NLP task, and one of the few applications of NeSy approaches to AM. In our opinion, there are several advantages in such an approach. From an analysis/interpretation perspective, logical rules play an active role not only during training but also at inference time, offering a means to investigate the behavior of the models. For example, we could easily measure compliance.

我们展示了 LTNs 在一个具有挑战性的 NLP 任务中的首次应用，以及 NeSy 方法在 AM 中的少数应用之一。在我们看来，这种方法有几个优点。从分析/解释的角度来看，逻辑规则在训练期间和推理时都发挥着积极作用，提供了一种调查模型行为的手段。例如，我们可以轻松地测量合规性。

From a user perspective, the definition of training rules and queries requires only a basic knowledge of FOL, which may contribute to reducing the divide between system architects and domain experts, who do not need to be also experts in machine learning, NeSy systems, or deep networks.

从用户的角度来看，训练规则和查询的定义只需要对 FOL 有基本的了解，这可能有助于缩小系统架构师与领域专家之间的差距，后者不需要同时成为机器学习、NeSy 系统或深度网络的专家。

From an architectural perspective, the decoupling between symbolic and neural components allows changing either of them without any direct impact on the other, except for the definition of key concepts such as the predicates/labels of the problem. Such a modularity may be highly beneficial in the context of AM, where one could use the same neural architecture with different corpora by expressing different symbolic rules. Indeed, the structural diversity of datasets and labeling schemes is a known issue in AM research, often leading to tailored solutions (Lippi and Torroni 2016).

从架构的角度来看，符号组件与神经组件之间的解耦允许在不直接影响彼此的情况下更改其中任何一个，除了定义关键概念，例如问题的谓词/标签。这种模块化在 AM 的背景下可能是非常有益的，在这种情况下，可以通过表达不同的符号规则使用相同的神经架构与不同的语料库。实际上，数据集和标注方案的结构多样性是 AM 研究中的一个已知问题，通常导致量身定制的解决方案 (Lippi 和 Torroni 2016)。

Performance-wise, the introduction of two symbolic rules did not negatively affect component classification performance and it increased link classification performance, while at the same time boosting robustness and compliance. It is worthwhile noticing that the networks used in our experiments are much simpler than state-of-the-art models, and obviously they do not achieve comparable performance, but we speculate that the impact of rules may hold even for more advanced models.

在性能方面，引入两个符号规则并未对组件分类性能产生负面影响，反而提高了链接分类性能，同时增强了鲁棒性和合规性。值得注意的是，我们实验中使用的网络比最先进的模型简单得多，显然它们无法达到可比的性能，但我们推测规则的影响可能在更先进的模型中仍然存在。

Dataset	Split	Approach	Classification		Agreement		Properties	
			Comp.	Link	Comp.	Link	Eq. 2	Eq. 3
Neoplasm	Validation	Data	83 - 84	42 - 41	77	66	98	100
		Data + Rules	84 - 85	44 - 43	81	71	100	100
Neoplasm	Test	Data	79 - 80	34 - 31	77	64	98	100
		Data + Rules	79 - 78	35 - 35	79	70	100	100
Glaucoma	Test	Data	82 - 82	45 - 43	75	66	99	100
		Data + Rules	81 - 82	47 - 45	75	71	100	100
Mixed	Test	Data	81 - 81	38 - 34	75	64	98	100
		Data + Rules	81 - 80	39 - 40	76	69	100	100

数据集	划分	方法	分类		协议		属性	
			竞争	链接	竞争	链接	方程 2	方程 3
新生物	验证	数据	83 - 84	42 - 41	77	66	98	100
		数据 + 规则	84 - 85	44 - 43	81	71	100	100
新生物	测试	数据	79 - 80	34 - 31	77	64	98	100
		数据 + 规则	79 - 78	35 - 35	79	70	100	100
青光眼	测试	数据	82 - 82	45 - 43	75	66	99	100
		数据 + 规则	81 - 82	47 - 45	75	71	100	100
混合	测试	数据	81 - 81	38 - 34	75	64	98	100
		数据 + 规则	81 - 80	39 - 40	76	69	100	100

Table 2: Results of NeSy AM on AbstRCT against the data-driven baseline. For component classification, we report both the result obtained by the MAJ approach (before the dash) and by the AVG approach (after the dash). Scores are reported as percentage values.

表 2:NeSy AM 在 AbstRCT 上相对于数据驱动基线的结果。对于组件分类，我们报告了 MAJ 方法 (破折号前) 的结果和 AVG 方法 (破折号后) 的结果。得分以百分比值报告。

On the down side, we shall remark that one major challenge for this kind of approaches is scalability to larger domains, and the fact that they are not specifically designed for NLP tasks, so their development is yet in its infancy.

不利的一面是，我们要指出这种方法面临的一个主要挑战是可扩展性，尤其是对更大领域的适应性，以及它们并不是专门为 NLP 任务设计的，因此它们的发展仍处于初期阶段。

As future work, we are considering the weighting of soft rules, so as to distinguish between rules expressing preferences (or theories) and those expressing constraints.

作为未来的工作，我们考虑对软规则进行加权，以区分表达偏好 (或理论) 的规则和表达约束的规则。

Once the scalability issue will be solved, we plan to experiment with larger corpora, more advanced embeddings, and deeper neural architectures. Moreover, it will be interesting to define rules that apply only to a subsets of entities. For example, in our benchmark, the argumentation graphs link only entities that belong to the same document, hence the collective classification may be performed document-wise, rather than dataset-wise, with the consequence that rules will be applied only between elements of the same document. Such a consequence may be a desired property or an unwanted drawback, according to the specific context. A collective classification on the whole corpus would be beneficial in applications where the argumentation spans across multiple domains and the aim is to find relations between components that belong to different documents. This approach suits contexts such as mining argumentation in social networks (Bosc, Cabrio, and Villata 2016) or retrieving arguments related to a specific topic (Ein-Dor et al. 2020).

一旦可扩展性问题得到解决，我们计划对更大规模的语料库、更加先进的嵌入和更深层的神经架构进行实验。此外，定义仅适用于某些实体子集的规则将是一个有趣的方向。例如，在我们的基准测试中，论证图仅链接属于同一文档的实体，因此集体分类可以按文档进行，而不是按数据集进行，这样的结果是规则仅在同一文档的元素之间应用。这样的结果可能是一个期望的属性，也可能是一个不希望出现的缺陷，具体取决于特定的上下文。在整个语料库上进行集体分类在论证跨越多个领域的应用中是有益的，目的是找到属于不同文档的组件之间的关系。这种方法适用于诸如在社交网络中挖掘论证 (Bosc, Cabrio, and Villata 2016) 或检索与特定主题相关的论据 (Ein-Dor et al. 2020) 等上下文。

Another direction regards the recognition of properties that are not explicit in the training data but can be defined through logical rules. In the context of our setting, it would be possible to define a predicate that represents a property without the need to provide any grounding example for it. This could be achieved by creating axioms that involve such a predicate and other grounded predicates, so as to train the neural network associated with the new predicate along with the ones for which the grounding is provided. This could allow the network to infer information regarding components or relations without labeled training data: for example, finding which claim is the major claim of a document, or which components agree with each other.

另一个方向涉及识别在训练数据中未明确表示但可以通过逻辑规则定义的属性。在我们设置的背景下，可以定义一个表示某种属性的谓词，而无需为其提供任何基础示例。这可以通过创建涉及该谓词和其他基础谓词的公理来实现，从而训练与新谓词相关的神经网络以及为其提供基础的谓词。这可以使网络在没有标记训练数据的情况下推断有关组件或关系的信息: 例如，找出哪个主张是文档的主要主张，或者哪些组件彼此一致。

References

参考文献

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Lev-enberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; and Zheng, X. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. CoRR abs/1603.04467. URL http://arxiv.org/abs/1603.04467.

Badreddine, S.; d'Avila Garcez, A.; Serafini, L.; and Spranger, M. 2020. Logic Tensor Networks. CoRR abs/2012.13635. URL https://arxiv.org/abs/2012.13635.

Bergmann, M. 2008. An introduction to many-valued and fuzzy logic: semantics, algebras, and derivation systems. Cambridge University Press.

Bosc, T.; Cabrio, E.; and Villata, S. 2016. Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media. In Baroni, P.; Gordon, T. F.; Schef-fler, T.; and Stede, M., eds., COMMA 2016, volume 287 of Frontiers in Artificial Intelligence and Applications, 21- 32. IOS Press. doi:10.3233/978-1-61499-686-6-21. URL https://doi.org/10.3233/978-1-61499-686-6-21.

Cabrio, E.; and Villata, S. 2018. Five Years of Argument Mining: a Data-driven Analysis. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, 5427-5433. International Joint Conferences on Artificial Intelligence Organization. doi:10. 24963/ijcai.2018/766. URL https://doi.org/10.24963/ijcai.2018/766.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; St. John, R.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; Strope, B.; and Kurzweil, R. 2018. Universal Sentence Encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 169-174. Brus-

sels, Belgium: Association for Computational Linguistics.

比利时塞尔斯: 计算语言学协会。

doi:10.18653/v1/D18-2029. URL https://www.aclweb.org/ anthology/D18-2029.

d'Avila Garcez, A. S.; Gori, M.; Lamb, L. C.; Serafini, L.; Spranger, M.; and Tran, S. N. 2019. Neural-symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. FLAP 6(4): 611-632. URL https://collegepublications.co.uk/ifcolog/?00033.

De Raedt, L.; Dumancic, S.; Manhaeve, R.; and Marra, G. 2020. From Statistical Relational to Neuro-Symbolic Artificial Intelligence. In Bessiere, C., ed., Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, 4943-4950. ijcai.org. doi:10. 24963/ijcai.2020/688. URL https://doi.org/10.24963/ijcai.2020/688.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1), 4171-4186. ACL.

Donadello, I.; Serafini, L.; and Garcez, A. D. 2017. Logic Tensor Networks for Semantic Image Interpretation. In IJCAI, 1596-1602. AAAI Press. ISBN 9780999241103. doi:10.24963/ijcai.2017/221. URL https://doi.org/10.24963/ ijcai.2017/221.

Dung, P. M. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence 77(2): 321-358. doi:10.1016/0004-3702(94)00041- X. URL http://www.sciencedirect.com/science/article/pii/ 000437029400041X.

Ein-Dor, L.; Shnarch, E.; Dankin, L.; Halfon, A.; Sznajder, B.; Gera, A.; Alzate, C.; Gleize, M.; Choshen, L.; Hou, Y.; Bilu, Y.; Aharonov, R.; and Slonim, N. 2020. Corpus Wide Argument Mining - A Working Solution. In AAAI, 7683-7691. AAAI Press. URL https://aaai.org/ojs/index.php/AAAI/article/view/6270.

Galassi, A.; Kersting, K.; Lippi, M.; Shao, X.; and Torroni, P. 2019. Neural-Symbolic Argumentation Mining: An Argument in Favor of Deep Learning and Reasoning. Frontiers in Big Data 2: 52. doi:10.3389/fdata.2019.00052. URL https://doi.org/10.3389/fdata.2019.00052.

Galassi, A.; Lippi, M.; and Torroni, P. 2021. Multi-Task Attentive Residual Networks for Argument Mining. CoRR abs/2102.12227. URL https://arxiv.org/abs/2102.12227.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. Deep Learning. MIT Press. URL http://www.deeplearningbook.org

Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; and Xing, E. 2016. Harnessing Deep Neural Networks with Logic Rules. In ACL (1), 2410-2420. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1228. URL https://www.aclweb.org/anthology/P16-1228.

Lippi, M.; and Frasconi, P. 2009. Prediction of protein β - residue contacts by Markov logic networks with grounding-specific weights. Bioinformatics 25(18): 2326-2333. ISSN

1367-4803. doi:10.1093/bioinformatics/btp421. URL https: //dx.doi.org/10.1093/bioinformatics/btp421.

Lippi, M.; and Torroni, P. 2016. Argumentation Mining: State of the Art and Emerging Trends. ACM Trans. Internet Technol. 16(2): 10:1-10:25. ISSN 1533-5399. doi:10.1145/ 2850417. URL http://doi.acm.org/10.1145/2850417.

Manhaeve, R.; Dumančić, S.; Kimmig, A.; Demeester, T.; and De Raedt, L. 2021. Neural probabilistic logic programming in DeepProbLog. Artificial Intelligence 298: 103504. ISSN 0004-3702. doi:https://doi.org/10.1016/j.artint.2021.103504.URL https://www.sciencedirect.com/ science/article/pii/S00043702210

Mayer, T.; Cabrio, E.; and Villata, S. 2020. Transformer-Based Argument Mining for Healthcare Applications. In Giacomo, G. D.; Catalá, A.; Dilkina, B.; Milano, M.; Barro, S.; Bugarín, A.; and Lang, J., eds., ECAI 2020 - 24th European Conference on Artificial Intelligence, volume 325 of Frontiers in

Artificial Intelligence and Applications, 2108- 2115. IOS Press. doi:10.3233/FAIA200334. URL https://doi.org/10.3233/FAIA200334.

Mayer, T.; Marro, S.; Cabrio, E.; and Villata, S. 2021. Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials. Artificial Intelligence in Medicine 102098. ISSN 0933-3657. doi:https: //doi.org/10.1016/j.artmed.2021.102098. URL https://www.sciencedirect.com/science/article/pii/S

Pacheco, M. L.; and Goldwasser, D. 2021. Modeling Content and Context with Deep Relational Learning. Trans. Assoc. Comput. Linguistics 9: 100-119. URL https://transacl.org/ojs/index.php/tacl/article/view/2381.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543. Association for Computational Linguistics. doi:10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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), 3982-3992. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19- 1410. URL https://www.aclweb.org/anthology/D19-1410.

Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Gallagher, B.; and Eliassi-Rad, T. 2008. Collective Classification in Network Data. AI Mag. 29(3): 93-106. doi:10.1609/ aimag.v29i3.2157. URL https://doi.org/10.1609/aimag.v29i3.2157.

Serafini, L.; and d'Avila Garcez, A. S. 2016. Learning and Reasoning with Logic Tensor Networks. In Adorni, G.; Cagnoni, S.; Gori, M.; and Maratea, M., eds., AI*IA 2016 Advances in Artificial Intelligence, 334-348. Cham: Springer International Publishing. ISBN 978-3-319-49130- 1.

Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In Burges, C. J. C.; Bot-tou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., Advances in Neural Information Processing Systems, volume 26, 926-934. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2013/file/ b337e84de8752b27eda3a12363109e80-Paper.pdf.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In NIPS, 5998-6008. Curran Associates, Inc.

Walton, D. 2009. Argumentation Theory: A Very Short Introduction. In Simari, G.; and Rahwan, I., eds., Argumentation in Artificial Intelligence, 1-22. Springer US. ISBN 978-0-387-98196-3. doi:10.1007/978-0-387-98197- 0 1. URL http://link.springer.com/chapter/10.1007/978-0- 387-98197-0_1.