IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 November 2023, accepted 26 November 2023, date of publication 30 November 2023,
date of current version 8 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3338189

RESEARCH ARTICLE

Toward Intent-Based Network Automation for Smart Environments: A Healthcare 4.0 Use Case

YOSRA NJAH[1], (Member, IEEE), ARIS LEIVADEAS[1], (Senior Member, IEEE), JOHN VIOLOS[1], AND MATTHIAS FALKNER[2]

[1]Department of Software and IT Engineering, École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada
[2]Cisco's SP Sales CTO Group, Cisco Systems Inc., Ottawa, ON K2K 3E8, Canada

Corresponding author: Aris Leivadeas (aris.leivadeas@etsmtl.ca)

This work was supported in part by the Cisco University Research Program Fund of Silicon Valley Community Foundation under Grant 2021-234759.

ABSTRACT Today's organizations have been embracing digital transformation to boost the quality of living within IoT-based smart-sustainable environments (e.g., healthcare, factories, vehicles, etc.). At the same time, augmenting the network infrastructure surface with billions of new devices accommodating myriad applications creates the need for network automation through different technologies, such as Software-Defined Networking (SDN), Network Function Virtualization (NFV), and Big Data Analytics (BDA). However, to devise an end-to-end self-driving and autonomous network, the manual configuration of network parameters and devices should be limited or even vanished. The recently emerged Intent-based Networking (IBN) paradigm introduces an additional building block enabling the network to adapt its settings automatically according to high-level user demands (intents) while hiding low-level details of the underlying infrastructure (e.g., configurations in millions of network devices). This paper initiates a deeper discussion regarding service automation over a Hospital 4.0 environment, from translating user requests to service profiling (unstructured intent refinement), deployment, and assurance. First, we discuss the design challenges of joining an intent-based framework as a convenient plane to an SDN-based platform. Following, we focus on an intelligent intent refinement system based on the Named Entity Recognition (NER) approach, an application of Natural Language Processing (NLP). This IBN-NER system deploys an extensible network policy model and the pre-trained Google's BERT (Bidirectional Encoder Representations from Transformers) algorithm, fine-tuned with a Healthcare 4.0 dataset. The proposed intent refinement framework is evaluated via extensive simulations with an incremental number of heterogeneous intents. Our simulation results show promising performance with only one epoch for all dataset sizes and all policy model entities tested. For example, with 5000 intents, our system provides the highest accuracy with 86%; meanwhile, the well-known benchmarks in the NER problem, namely BiLSTM-CRF, BiLSTM, and LSTM, with ten epochs, provide 57%, 31%, and 26%, respectively.

INDEX TERMS Healthcare 4.0, intent-based networking, network automation, intent refinement, service policy mapping, named entity recognition.

I. INTRODUCTION

Smart environments, particularly the ones related to Industry 4.0, are fundamentally changing services and production worlds. For example, the health domain and its whole ecosystem is moving towards healthcare 4.0 by embracing key enabling technologies, such as the Internet of Things (IoT), Cyber-Physical Systems (CPS), and data analysis, among others. With this transformation, global spending on healthcare is expected to increase to $18.28 trillion worldwide by 2040 to build up agility and effectiveness in all directions [1], [2].

At the same time, the emergence of network softwarization and enhanced device programmability and monitoring offer the pedestal for autonomic management. Thus, an autonomous network, or self-driving network, has become a strong worldwide interest for service providers to fully automate massive heterogeneous network infrastructures

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Arafatur Rahman.

[3]. This automation should cover all aspects of network design (i.e., configuration, optimization, healing, etc.) while adapting to an ever-increasing infrastructure scale. The latter is of utmost importance for IoT and smart environments (e.g., hospitals, factories, campuses, homes, and offices), where the number of devices and applications disfavors human intervention and manual configuration. To this end, various industrial companies, such as Cisco [4], IBM [5], Juniper [6], etc., are already making efforts to design end-to-end autonomous networks to ensure efficient, reliable, and secure management of massive smart environments.

Automating network management operations is conducted by different technologies, such as Software-Defined Networking (SDN), Network Function Virtualization (NFV), and Artificial Intelligence (AI) models. Intent-Based Networking (IBN) is a novel paradigm that can be proved to be a critical component towards fulfilling the arrival of a next-generation management system, which is the zero-touch service and network management (ZSM) [7], [8].

IBN allows users to express in a declarative way (e.g., human-natural language) general high-level interests or concerns (e.g., service, performance, data storage, etc.) and in an autonomous way to configure the underlying network infrastructure according to users' intents. It is beyond a game changer, where using advanced data analytics mechanisms, intents are translated into network policies (i.e., executable scripts) and implemented with the power of IT automation into the network to enhance the agility of IT infrastructure in every possible aspect. In particular, IBN creates a closed-loop automation system, including the processes of intent expression, intent refinement, intent activation, and intent assurance, that replaces error-prone and manual network configurations with intelligent and advanced softwarized mechanisms [9].

Even though IBN is a promising solution and despite the efforts of major international standards organizations (e.g., IETF [10]) and industries (e.g., Cisco [11] and Huawei [12]) to define a proper reference model, almost every IBN component is still in its infancy. Yet, IBN is critical for developing future applications over a completely autonomous network, particularly with the new emerging network scenarios coming from different technologies, such as smart environments, IoT, CPS, and 6G networks.

One of the critical initial components in designing IBN is the intent expression interface. This interface will allow any type of user (from a simple end-user to a network administrator) to express what they want from the network. There are two typical sorts to express an intent, either through a human-readable form (i.e., a Graphical User Interface) to define the type of service/application along with some high-level and qualitative requirements (e.g., Quality of Service, security, etc.), or via natural language. However, the latter fashion requests more sophisticated techniques based on Natural Language Processing (NLP) to infer the intent [13].

In this paper, we consider the intent-based natural language input. Additionally, we introduce a high-level reference intent-based networking architecture that can leverage a fully programmable network infrastructure. Its design is based on the requirements of IoT-based smart environments, e.g., Healthcare 4.0 as a use case, to meet an end-to-end self-driving network. In particular, we propose an intelligent intent refinement engine to automatically translate user-defined intents ("what to do") into network policies ("how to do it"). It is noteworthy that intent refinement involves the first two steps of a typical IBN platform (i.e., intent expression and policy translation) that enable the map of intents from a declarative language to a machine-readable policy. Most of the existing research proposals for IBN flunk against the validation of intent translation, and they lack in performance evaluation (e.g., [13] and [14]). Thus, in this paper, a deeper discussion regarding IBN-based service automation is initiated, from the inference of user intents and service profiling to their deployment and assurance, while focusing especially on the intent refinement steps of the closed loop. Moreover, we address the IBN-NLP challenge using the Named Entity Recognition (NER) approach, utilizing artificial intelligence techniques and pre-trained models for effective resolution [15]. This approach surpasses traditional methods, such as knowledge graphs, by leveraging deep contextual understanding and adaptability to diverse datasets, resulting in superior precision, efficiency, scalability, and versatility for handling heterogeneous intents within a large-scale smart environment.

The contributions of this paper are summarized as follows:

- An IBN-enabled architecture applied to a hospital 4.0 use case is introduced for the first time. The architecture combines the benefits of a fully programmable SDN-based network and IBN, while considering realistic intents that stem from various healthcare aspects such as patient care, health professionals, and resources. The latter builds a dataset with a massive number of heterogeneous intents crucial to the efficiency of the intent refinement engine.
- An NLP-based intelligent intent refinement is designed following a fine-grained Named-Entity Recognition (NER) approach, a crucial form of NLP, to extract and map keywords (called named-entities) from an unstructured intent to a network policy model. This IBN-NER problem is addressed by using, for the first time, to the best of our knowledge, Google's emerging NLP model called Bidirectional Encoder Representations from Transformers (BERT) [16]. The model is fine-tuned on a large corpus of labeled entities with 40000 intents specific to smart healthcare network tasks.
- An exhaustive simulation is conducted considering various dataset sizes of simultaneous heterogeneous intents, and the proposed intent refinement system is evaluated in terms of processing time and accuracy with

global and granular measures. Compared to well-known benchmarks in the NER domain (i.e., LSTM, Bi-LSTM, and BiLSTM-CRF [13]), our simulation results show promising performance with only one epoch, different intent dataset sizes, and policy's entities.

The remainder of this paper is organized as follows. Section II discusses relevant related works in intent-based networking research. Section III presents an IBN-enabled Healthcare 4.0 use case over a generic fully-programmable network architecture. Section IV introduces the deployment of the intelligent IBN refinement system. Section V assesses the performance of the proposed approach. Finally, the paper concludes and discusses future works in Section VI.

II. RELATED WORK

This section highlights relevant existing works from the perspective of intent-based networking and explores how intent-based systems receive users' intents as input. Additionally, a brief discussion is provided regarding network automation in the healthcare domain.

A. INTENT-BASED NETWORK PARAMETER INPUT

Most of the existing approaches leverage and extend the capabilities of SDN by enhancing its northbound interface to accept the intents. For example, in [14], the authors presented an intent-based network for data forwarding in software-defined vehicular edge computing. The intent processing is performed using an ONOS intent controller and an ML approach (convolutional neural network) that categorizes incoming intents into three priority traffic classes routed through different paths. Similarly, in [17], the IBN-enabled ONOS interface was used to set up different network configurations to enhance 5G service management. Furthermore, Sanvito et al. [18] extended the ONOS intent framework to compile multiple intents simultaneously, which were lately translated into network routes. Nevertheless, current SDN-based IBN proposals enable to submit intents only using the command line interface, the RESTful interface, or a native Python API that not all users within an intelligent environment (i.e., doctors in a hospital) are able to use.

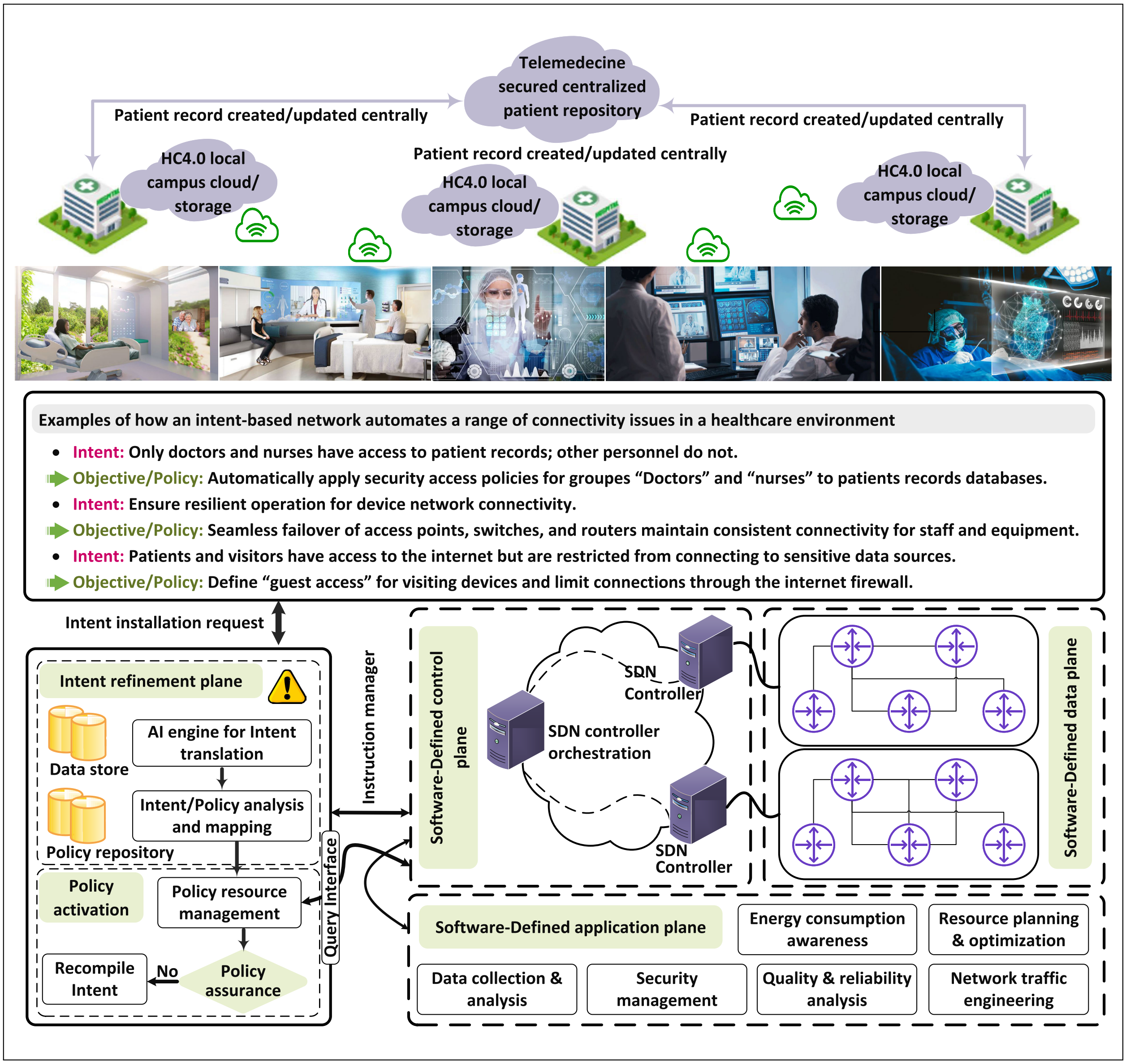
Instead of using an SDN interface, the intent could also be submitted through more friendly Graphical User Interfaces (GUI) in the form of drop-down menus. For instance, the authors in [19] proposed an IBN framework in the context of 5G network slicing. Specifically, the users can express network intents via a GUI with different fields corresponding to the deployment, management, and monitoring of network resources [20]. Even though this GUI-based approach enables easy intent generation and policy mapping, it is restrictive in terms of options and does not allow the end-user to provide any other input with different levels of detail or granularity.

B. INTENT-BASED NATURAL LANGUAGE INPUT

Although the above approaches make steps towards abstracting the network requirements, they are quite rigid and usually targeting network specialists. Thus, lately, a natural language approach started to emerge as a means to express intent. For example, the authors, in [21], utilized Amazon voice-assisted technologies and a Latent Dirichlet Allocation (LDA) NLP approach to perform the intent refinement. The framework includes intents related to five network scenarios, namely authentication, network security, performance, access control, and self-healing. However, with the IBN problem, intents are presented with simple and short sentences; meanwhile, the LDA approach is generally unsuitable for document topic modeling if the dataset is too small, documents' lengths are too short, or there are too many topics within the dataset. Bensalem et al. [22], [23] presented intent-based networking in information and communications technology (ICT) supply chain networks. They extracted information from unstructured intents and stored them in JSON files inferred with an ML recommender that estimates the computational performance of different ML-based techniques (Singular Value Decomposition and Stochastic Gradient Descent). Similarly, in [24], the authors proposed intent-based solutions for automatic network orchestration through the application of Graph Neural Networks (GNN) and Long Short-Term Memory (LSTM) techniques. However, in [22] and [24], the authors did not provide extensive analysis of the deployed intent datasets, IBN-engineering systems, and their corresponding performances. In [13], Ouyang et al. reviewed the key-enabling technologies of the IBN while focusing on intent refinement schemes differentiated according to target users, input methods, and refinement approaches. Furthermore, they designed an intent refinement system based on NLP and deep learning techniques, showing the outperformance of the Bidirectional-LSTM and Coordination Random Field (BiLSTM-CRF) compared to the LSTM-CRF model. However, they did not provide details about the use cases and policy models, while the system performance lacked meticulous evaluation.

C. NETWORK AUTOMATION IN HEALTHCARE AND IoT

Until now, almost all of the existing IBN approaches cover generic network scenarios and disregard the domain that an IBN system will be applied. In the context of IoT, there are only a few preliminary works that study the use of IBN (i.e., [25], [26]) without considering any use case scenario or a natural language expressed intent. Regarding healthcare, there have recently been some efforts towards automating its network infrastructure through the integration of SDN in order to facilitate network management and reduce manual configuration [27]. However, the authors provide a high-level architecture and do not consider the integration of IBN, SDN, and healthcare. It is thus clear that IBN is still in its first development steps and not yet considered in the automation of smart environments, particularly healthcare. Accordingly, and to the best of our knowledge, in this paper, we present a first-of-its-kind architecture and intent refinement approach targeting a healthcare use case. Additionally, in contrast with


Telemedecine secured centralized patient repository
Patient record created/updated centrally
Patient record created/updated centrally
Patient record created/updated centrally
HC4.0 local campus cloud/ storage
HC4.0 local campus cloud/ storage
HC4.0 local campus cloud/ storage
Intent installation request
Intent refinement plane
Data store
AI engine for Intent translation
Intent/Policy analysis and mapping
Policy repository
Policy activation
Policy resource management
Recompile Intent
No
Policy assurance
Query Interface
Instruction manager
Software-Defined control plane
SDN Controller
SDN controller orchestration
SDN Controller
Software-Defined data plane
Software-Defined application plane
Energy consumption awareness
Resource planning & optimization
Data collection & analysis
Security management
Quality & reliability analysis
Network traffic engineering

Examples of how an intent-based network automates a range of connectivity issues in a healthcare environment

- Intent: Only doctors and nurses have access to patient records; other personnel do not.
- Objective/Policy: Automatically apply security access policies for groupes “Doctors” and “nurses” to patients records databases.
- Intent: Ensure resilient operation for device network connectivity.
- Objective/Policy: Seamless failover of access points, switches, and routers maintain consistent connectivity for staff and equipment.
- Intent: Patients and visitors have access to the internet but are restricted from connecting to sensitive data sources.
- Objective/Policy: Define “guest access” for visiting devices and limit connections through the internet firewall.

FIGURE 1. Intent-based network automation architecture over a Hospital 4.0 as a use case for smart environments.

other existing NLP approaches, we propose for the first time the use of BERT to facilitate intent capturing and translation.

III. INTENT-BASED NETWORK AUTOMATION ARCHITECTURE OVER A HOSPITAL 4.0

In this section, an IBN-enabled architecture is proposed over a Hospital 4.0 as a real-life use case for smart environments. This architecture can facilitate a wide range of intents, enabling the assurance of heterogeneous services within the digital healthcare ecosystem and supporting generic network management operations. Furthermore, we review the persistent need to integrate an intent-based approach into a fully-programmable architecture to achieve an end-to-end self-driving network.

Indeed, the healthcare ecosystem, especially in the post-COVID-19 era, is undergoing an accelerated digital transformation towards Healthcare 4.0 [28]. This transformation offers a diverse spectrum of high-quality healthcare services, including, for example: remote patient monitoring through the use of IoT devices for tracking patient health [29]; telemedicine to enable remote healthcare services and consultations [30]; Electronic Health Records (EHR) for efficient storage and management of patient’s medical records [31]; healthcare data analytics (e.g., medical imaging) for data-driven decision-making [32]; and Health Information Exchange (HIE) for secure sharing of medical data among healthcare providers and institutions [33], [34]. These are just a few examples of the multifaceted changes taking place in healthcare and achieved through real-time monitoring of patients, healthcare experts/workers, and all hardware and software resources. As shown in Table 1, all Healthcare 4.0 applications are characterized by specific requirements, such as security, QoS provisioning, and real-time access. In the meantime, certain applications also need

TABLE 1. Example of Healthcare 4.0 application scenarios with their corresponding requirements.

Healthcare 4.0 application (Ex. of category)		App. requirements
Patient care	- Hospital patient care - In patient care - Patient/Public people care - Personalized healthcare - Home care	- Security - Privacy - Real-time access - Mobility
Healthcare experts/workers	- Management - Scheduling - Reducing "burnout" - Access to resources - Collaboration - Remote access	- Security - Real-time access - Mobility - Remote access - Collaboration
Healthcare resources	- Availability - Allocation - Management - Optimization - Automation - Fault management	- Security - Real-time access - Mobility - Interoperability - Fault management

```
Algorithm 1 . From Intent Refinement to Assurance
Input: $I_L$ ▷ Intent list
Input: $P_M$ ▷ Policy model for various networking aspects
Input: $P_D$ ▷ Network declarative policy (Initially is empty)
Input: $R_M$ ▷ SDN-based available network resource
Output: $Scripts$ ▷ Actions related to network resource configuration
Intent Refinement to Assurance ($I_L, P_M, P_D, R_M$)
1: while all intents $I \in I_L$ have not been processed do
2:     Mapping $I$ to $P_D$ using AI engines and $P_M$
3:     Real-time $R_M$ updating ▷ Continuous SDN topology
       monitoring (resources and demands)
4:     Assess $R_M$ to infer if $P_D$ could be fulfilled
5:     if $P_D$ could be fulfilled then
           ▷ Configure network resources with $P_D$ to satisfy $I$
               $Scripts \leftarrow P_D$
6:     else
           ▷ Report $P_D$ as violated to plan resolution based
             on updated $R_M$ and the intent $I$ flexibility.
7: Return $Scripts$
End of Intent Refinement to Assurance
```

additional requirements, such as privacy for patient records, mobility and remote access for medical experts, and fault management for resource control [2]. Hence, the designed intent dataset is based on a list of network scenarios, including service performance, cloud storage, security, authentication, access control, and self-healing.

Figure 1 shows an intent-based network automation architecture over a Hospital 4.0 (as a use case) while considering the dynamic and complex natures of such an ecosystem. The first layer of this prototype displays the healthcare services deployed through a combination of communications technologies and medical expertise. It connects a wide variety of heterogeneous facilities [28], such as healthcare CPSs, IoT medical devices and robots, clinical appliances, device gateways, and billing systems, all assisted with an intent-based input system utilized by simple users and network administrators that only indicate what needs to be done with human-natural language.

On a lower level, the IBN automation layers are designed with their corresponding intelligent systems for heterogeneous intent refinement and policy activation and assurance. The intent-based platform is designed as a convenient northbound layer to the typical SDN-based paradigm, including its control, forwarding, and application planes. This architecture enables us to reach an end-to-end self-driving network by deploying various network functionalities, such as data analytics, QoS provisioning, resource optimization, energy awareness, and self-defense. It is also noteworthy that among the key enabling technologies for this digital infrastructure, the proposed architecture supports virtualization, high-performance connectivity, and Edge/Cloud software, enabling reliable and flexible data management between experts [35], [36], [37], [38]. Let us consider the following example as an IBN-based scenario:

Remote diagnosis and robotic-assisted surgery: To perform surgery for a patient who is located in Montreal (Canada) and unable to travel to the United States, an expert surgeon in Cambridge may ask "*We need a robust connectivity in room 1 for remote robotic-assisted surgery*." Hence, the IBN platform should map this intent to a network policy related to slicing, where the dedicated network slices ensure high bandwidth, ultra-high speeds, and almost zero latency to enable real-time remote surgery. In more detail, as shown in Figure 1 and summed in Algorithm 1, the intent refinement plane receives this intent as a human utterance. It infers it to define in a granular manner how best to implement this request in the network infrastructure. An *AI-based translation* engine extracts and analyses the intent's key elements (e.g., words) that will be later mapped via an *intent-policy* model P_M to a simplified structured intent (declarative network policy P_D), which is used as *actionable scripts* for implementation by the *Policy resource management* module, Algorithm 1, line 2. In the meantime, the SDN controller continuously monitors network status R_M and makes assertions to verify any conflicts between the policy (refined intent configuration) and the actual network state (Algorithm 1, lines 3-5). Policy's activation is optimized and rigorously evaluated for assurance, including potential trade-offs and conflicts of interest. Hence, the platform achieves the IBN-based closed-loop system, consisting of the three main deployment steps: Refinement, Activation, and Assurance.

IV. INTENT REFINEMENT MODEL

This section discusses the proposed intelligent IBN refinement system. In particular, we design a fine-grained NER approach, a crucial form of NLP, to i) extract keywords (called named-entities) defining application requirements and their network behavior from user utterances (unstructured intents), ii) convert them to specific configurations via an extensible policy model to be set up through the management schema on the available network resources.

A. APPLYING NER TO THE IBN-NLP PROBLEM

To identify the IBN-related information, a policy model is introduced, which is represented as a 5-tuple of entity labels as follows: ⟨*user, application goal/utility, network action, target equipment, timeframe*⟩. This policy model, as shown in Figure 2, enables compiling an unstructured user intent into a structured one. Hence, the "user" entity presents the intent definers at various abstraction layers. It could be technical users with specialized knowledge of networks (e.g., network administrators, operators, service providers, etc.), or non-technical users without specialized knowledge of networks (doctors, nurses, finance department, etc.). The field related to the "application goal/utility" supports users' demands, services, and objectives (e.g., remote diagnosis, data storage and accessing, etc.), while the "network action" entity presents the service and network management requirements to fulfill the intent. The model is flexible enough to support a wide variety of network and service management actions (e.g., QoS provisioning, orchestration, authentication and privacy, self-healing, and defense). On the other hand, the "target" field represents the targeted infrastructure (e.g., domain, hospital campus, region, provider, etc.) to deploy the user's intent. Finally, the "timeframe" entity defines the period during which the required service is scheduled to occur. It is noteworthy that all entities should be displayed within a user's utterance. Otherwise, a default setup will be applied. For example, if the expert surgeon asks, *"We need a robust connectivity in room 1 for remote diagnosis."*, a possible entity identification would be: {user: 'expert surgeon'}, {goal: videoconferencing with 'remote diagnosis'}, {network action: QoS provisioning with 'robust connectivity'}, {target: slice 1 with 'room 1'}, {timeframes: now with 'default selection'}. Table 2 shows the statistics of our dataset, which contains 40000 intents annotated manually with 1671 user entities, 1563 goal/utility entities, 3736 network action entities, 1981 target equipment entities, and 2029 timeframe entities.

B. APPLYING BERT TO THE IBN-NER PROBLEM

We address this IBN-NER problem using Google's BERT model, an open source and one of the most successful machine learning neural network-based frameworks for performing NLP tasks [16]. Figure 3 shows the deployed BERT-based architecture.

BERT is based on the transfer learning paradigm, where two separate stages are adopted: pre-training and fine-tuning. It is an encoder-only model, which includes several encoder blocks used initially for the pre-training phase. First, each intent is split into a sequence of individual tokens (entity-level tokenization) and padded to a maximum sequence length (e.g., 512 tokens per sequence with BERT). Each tokenized intent is annotated with the special tokens '[CLS]' and '[SEP]' to indicate each sequence's start and end, respectively. Following, each token in the sequence is projected into embeddings (sequential numeric vectors encoding semantic information) and fed as input to the encoder block. Before passing input embeddings to the transformer encoder, it adds the positional embeddings (position vectors) to process natural language sequences.

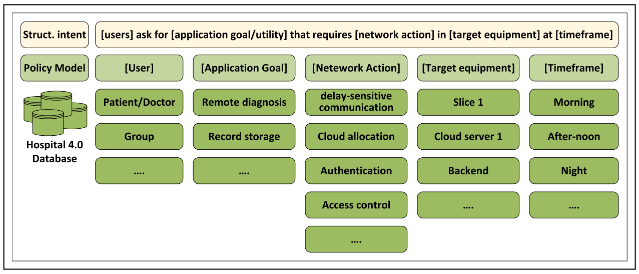


FIGURE 2. Policy model based on structured-intent.

TABLE 2. Intent dataset and policy entities.

Intent dataset	Entities				
	User	Goal	Action	Equipment	Timeframe
40000	1671	1563	3736	1981	2029

The first layer in the encoder block is a multi-head self-attention layer that determines multiple relationships between the embeddings by calculating a weighted average of all these encoded input vectors with a linear combination that pays attention to the similarity score between embeddings. Formally, the attention creates three main vectors: queries Q, keys K, and values V to extract feature representations, while the output vector is a weighted sum of V, where the weight specified for each value is identified by the dot products of the query with all the keys, which can be defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{n}}\right)V, \tag{1}$$

where n denotes the dimension of K and V and K^T is the transpose of the K keys. The multi-head self-attention linearly processes Q, K, and V multiple times via different weight matrices, i.e., W^Q, W^K and W^V respectively. Its output is another linear transformation via learnable parameters W^O of the concatenation of heads. Hence, the multi-head self-attention can be defined as:

$$MultiHead(Q, K, V) = \left[head_1, \ldots, head_h\right]W^O, \tag{2}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{3}$$

where h denotes the total number of heads. Then, the second layer in the encoder block is the feed-forward neural network module, which makes a higher representation for the attention outputs by adding non-linearity in the system, so it can learn complex relationships between

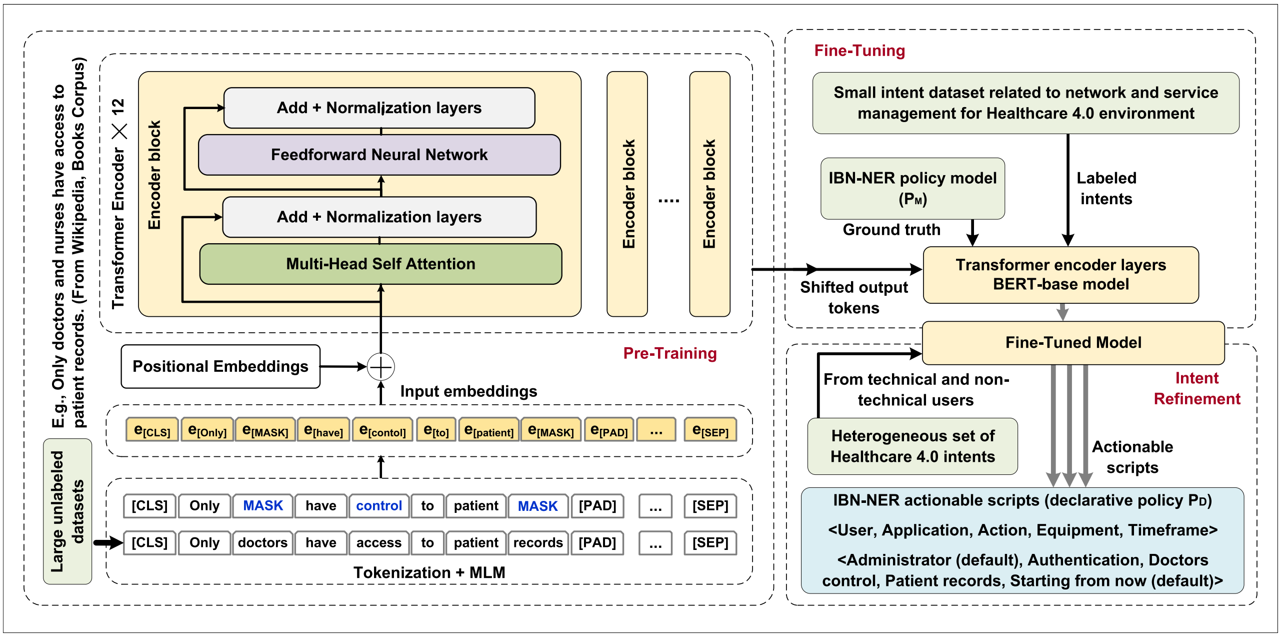

E.g., Only doctors and nurses have access to patient records. (From Wikipedia, Books Corpus)
Large unlabeled datasets
Transformer Encoder × 12
Encoder block
Add + Normalization layers
Feedforward Neural Network
Add + Normalization layers
Multi-Head Self Attention
Encoder block
....
Encoder block
Positional Embeddings
Input embeddings
Pre-Training
e[CLS] e[Only] e[MASK] e[have] e[contol] e[to] e[patient] e[MASK] e[PAD] ... e[SEP]
[CLS] Only MASK have control to patient MASK [PAD] ... [SEP]
[CLS] Only doctors have access to patient records [PAD] ... [SEP]
Tokenization + MLM
Fine-Tuning
Small intent dataset related to network and service management for Healthcare 4.0 environment
IBN-NER policy model (PM)
Ground truth
Labeled intents
Shifted output tokens
Transformer encoder layers BERT-base model
Fine-Tuned Model
From technical and non-technical users
Heterogeneous set of Healthcare 4.0 intents
Intent Refinement
Actionable scripts
IBN-NER actionable scripts (declarative policy PD)
<User, Application, Action, Equipment, Timeframe>
<Administrator (default), Authentication, Doctors control, Patient records, Starting from now (default)>

FIGURE 3. BERT-based named-entity recognition architecture for intent-based network refinement.

embeddings. At the production of these two layers, the encoder block applies "add and norm" layers to make normalization and provide the hidden state vectors to the input of each next layer, respectively. Thus, after going through 12 encoder blocks (BERT-base model), the system provides output embeddings that can be contextualized, where the same token will have different output embeddings depending on its surrounding terms. These output embeddings are also obtained due to the deployment of the Masked Language Model (MLM), one of BERT's significant innovations.

The MLM addresses the problem of 'see itself' that trains the model to predict any token based on the sequence's context. It is done by randomly selecting a percentage of the input tokens (e.g., 15%) during training and handling them with the masking operations. Hence, of that percentage (i.e., 15%), 80% are replaced with a special [MASK] token, 10% are randomly replaced by other arbitrary tokens, and the remaining 10% of the chosen tokens are kept unchanged. Specifically, as shown in the bottom left of Figure,3, given an input intent I of n tokens, the MLM replaces the i^{th} token with a special $[MASK]$ token, resulting in a new sequence X given by $X = \left[x_1, , \ldots, , x_{i-1}, [MASK], , x_{i+1}, , \ldots, , x_n\right]$.

The objective of MLM is to predict the original token, x_i that is masked, given its surrounding context, while learning a probabilistic model p_θ that minimizes the following loss function:

$$Loss_{MLM} = \sum_{x_i \in X_{mask}} -log\left(p_\theta(x_i)\middle|X\right). \quad (4)$$

Once BERT is trained on a vast unlabeled corpus (Books consisting of 800M words and English Wikipedia consisting of 2500M words), we fine-tune it with a labeled intent dataset related to service and network management over an intelligent healthcare environment, as discussed in the previous section. Hence, another major innovation of BERT is its ability to be fine-tuned for a specific domain, where the pre-trained MLM could be further trained for particular downstream NLP tasks using a small amount of labeled data. For example, consider a prediction task where $y \in \gamma$ is the target variable, e.g., the healthcare label. Fine-tuning uses gradient descent to modify the pre-trained parameters θ and learn a new set of parameters Φ in order to minimize:

$$Loss_{finetuning} = \sum_{X \in D} -log\left(p_{\theta,\Phi}(y\middle|X)\right), \quad (5)$$

where $p(y|x)$ is the ground-truth distribution and D is the data distribution of the downstream task. This fine-tuning operation allows BERT's parameters to be adjusted to better suit the healthcare management network and rarely requires training the model from scratch, which is tremendous, particularly because the healthcare data are limited, sensitive, and not publicly available for privacy reasons.

V. SIMULATION AND RESULT ANALYSIS

This section presents the simulation settings of the experiments and assesses the performance of the proposed framework against the state-of-the-art IBN-NLP algorithms from the pertinent literature.

A. SIMULATION SETTINGS

All experiments were conducted on a PC running an Intel Core i7 CPU @3.60GHz with 32.0 GB of memory.

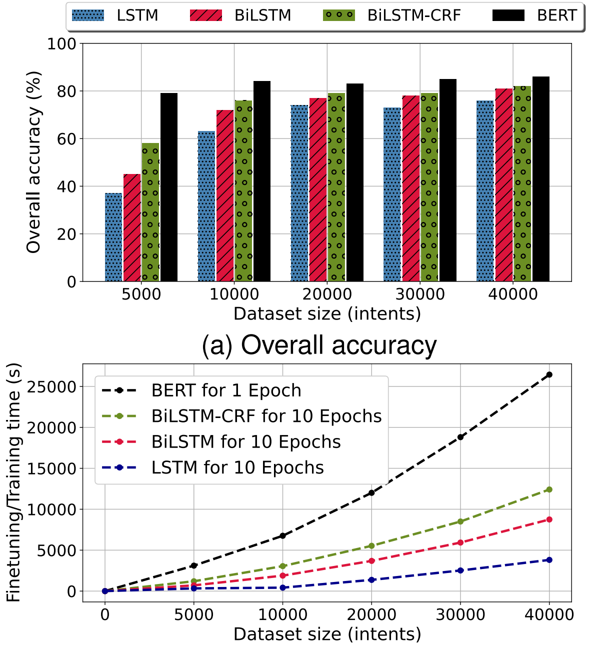
TABLE 3. Overall performance (BERT within 1-3 epochs and the benchmarks within 10-30 epochs.)

Dataset	Schemes	LSTM				BiLSTM				BiLSTM-CRF				BERT				
	Number of epochs	Time (s)	Precion	Recall	F1-score	Time (s)	Precion	Recall	F1-score	Time (s)	Precion	Recall	F1-score	Number of epochs	Time (s)	Precion	Recall	F1-score
5000	10 Epochs	330	0.37	0.26	0.20	450	0.45	0.31	0.36	990	0.58	0.56	0.56	1 Epoch	3210	0.79	0.80	0.79
10000		420	0.63	0.64	0.63	690	0.72	0.72	0.72	2050	0.76	0.74	0.75		6498	0.84	0.80	0.81
20000		1470	0.74	0.76	0.75	2830	0.77	0.74	0.76	4260	0.79	0.78	0.78		12405	0.83	0.82	0.82
30000		2820	0.73	0.77	0.75	4190	0.78	0.79	0.78	7950	0.80	0.79	0.79		20115	0.85	0.83	0.84
40000		3610	0.76	0.78	0.77	7500	0.81	0.81	0.81	13610	0.82	0.81	0.81		26443	0.86	0.85	0.85
5000	20 Epochs	441	0.61	0.59	0.59	1170	0.67	0.65	0.66	2428	0.69	0.67	0.68	2 Epochs	6714	0.79	0.80	0.79
10000		820	0.72	0.74	0.73	3880	0.76	0.75	0.75	4284	0.75	0.73	0.74		12917	0.82	0.82	0.82
20000		3220	0.73	0.75	0.74	5180	0.78	0.76	0.77	11410	0.77	0.75	0.76		24721	0.83	0.82	0.82
30000		5740	0.73	0.76	0.74	10660	0.79	0.77	0.78	16530	0.79	0.78	0.78		39766	0.86	0.85	0.85
40000		6900	0.82	0.81	0.81	18420	0.79	0.78	0.78	24360	0.81	0.79	0.80		52374	0.87	0.86	0.86
5000	30 Epochs	930	0.67	0.68	0.67	2680	0.69	0.68	0.68	5420	0.70	0.66	0.68	3 Epochs	9524	0.82	0.78	0.79
10000		2130	0.70	0.73	0.71	3660	0.74	0.73	0.74	7950	0.74	0.72	0.73		19484	0.84	0.83	0.82
20000		4650	0.71	0.75	0.73	8260	0.76	0.76	0.76	14230	0.77	0.75	0.76		37561	0.85	0.82	0.83
30000		9030	0.72	0.76	0.74	11700	0.78	0.78	0.78	19440	0.79	0.78	0.78		58558	0.86	0.85	0.85
40000		13600	0.73	0.76	0.75	20640	0.78	0.78	0.78	36570	0.79	0.79	0.79		79344	0.86	0.85	0.85

We implement the proposed BERT-based IBN scheme using the PyTorch-Transformers library [41]. Meanwhile, LSTM-based NER benchmarks described below are implemented using Python's Scikit-Learn [42], Tensorflow [43], and Keras [44] libraries. Regarding the dataset, as mentioned in Section IV, we trained the models using 40 000 labeled intents with more than 1000000 different words. Table 2 shows the statistics of the dataset used, which was split into two subsets, one containing 80% of the instances for training and another containing 20% for testing.

We adopt the BERT-base architecture, which includes 12 transformer blocks, 12 attention layers, and 110 million parameters. The BERT-based model is fine-tuned for a maximum of 3 epochs only while randomly masking 15% of all tokens within each sequence. As benchmarks, the LSTM, BiLSTM, and BiLSTM-CRF models are used since they are considered in the IBN literature as the most successful algorithms for NER tasks [13]. Specifically, the BiLSTM algorithm is built as a combination of two LSTM compound layers for forward and backward context inputs to capture the entire essence of the intent. Meanwhile, the BiLSTM-CRF model adds a Conditional Random Field layer to the BiLSTM scheme to infer the inter-dependency of each location, particularly with the neighboring labels in a sequence. The 5-fold cross-validation approach was used to determine the appropriate hyperparameter combination. Furthermore, the LSTM-based models were built using three layers while setting the number of hidden units to 75 and the embedding vector size to 128. Finally, the Adam optimizer with a learning rate of 0.005 was used, along with the Softmax as an activation function, and the sparse categorical cross-entropy (for LSTM and BiLSTM) and Keras-Contrib (for CRF-layer) as loss functions to optimize the output layer.

To evaluate the BERT-based intent refinement platform, we examine the performance from different aspects, including the training time and the refinement accuracy (precision, recall, and F1-score), while addressing five sizes of datasets: 5000, 10000, 20000, 30000, and 40000 entries. To observe the convergence to optimal solutions, a range of [10 to 30] epochs, with a step of 10, was examined for the three LSTM-based models. In the following and for better visualization, the results are outlined with gradient colors (blue and yellow) from 0% to 100% according to the corresponding metric. For example, the darkest blue (Table 3) and yellow (Table 4) cells highlight poor prediction results, while the lightest cells represent the highly accurate results.



(a) Overall accuracy

(b) Fine-tuning/Training time

FIGURE 4. Overall performance of BERT with 1 epoch and the benchmarks with 10 epochs.

B. RESULTS AND OBSERVATIONS

Table 3 presents the overall performance of all schemes (BERT within 1-3 epochs and the benchmarks within

TABLE 4. Policy model's entity inspection with 1 epoch for BERT and 10 epoch for the benchmarks.

Models		LSTM			BiLSTM			BiLSTM-CRF			BERT		
DS	Entity	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
5000	User	0.14	0.27	0.18	0.31	0.31	0.31	0.44	0.45	0.45	0.96	0.91	0.93
	Application utility/goal	0.29	0.61	0.40	0.29	0.66	041	0.58	0.76	0.66	0.59	0.90	0.72
	Network action	0.41	0.02	0.03	0.78	0.09	0.15	0.88	0.65	0.75	0.91	0.71	0.79
	Target equipment	0.01	0.00	0.00	0.57	0.01	0.02	0.33	0.20	0.25	0.68	0.60	0.64
	Timeframe	0.72	0.14	0.23	0.78	0.18	0.29	0.72	0.61	0.66	1.00	0.82	0.90
10000	User	0.43	0.47	0.45	0.62	0.61	0.61	0.66	0.63	0.64	0.93	0.93	0.93
	Application utility/goal	0.66	0.79	0.72	0.73	0.80	0.76	0.77	0.82	0.80	0.70	0.86	0.77
	Network action	0.88	0.82	0.85	0.91	0.88	0.89	0.91	0.89	0.90	0.82	0.86	0.84
	Target equipment	0.37	0.36	0.37	0.49	0.49	0.49	0.67	0.55	0.61	0.86	0.49	0.62
	Timeframe	0.77	0.65	0.70	0.84	0.74	0.79	0.83	0.77	0.80	0.92	0.58	0.71
20000	User	0.70	0.69	0.69	0.67	0.61	0.64	0.70	0.65	0.67	0.92	0.92	0.92
	Application utility/goal	0.76	0.84	0.80	0.80	0.82	0.81	0.81	0.85	0.83	0.66	0.92	0.77
	Network action	0.94	0.89	0.92	0.94	0.90	0.92	0.95	0.89	0.92	0.93	0.88	0.90
	Target equipment	0.49	0.55	0.52	0.61	0.57	0.59	0.65	0.61	0.63	0.77	0.54	0.63
	Timeframe	0.77	0.77	0.77	0.84	0.77	0.80	0.84	0.82	0.83	0.88	0.76	0.82
30000	User	0.68	0.69	0.68	0.68	0.69	0.68	0.71	0.65	0.68	0.90	0.93	0.91
	Application utility/goal	0.76	0.79	0.78	0.82	0.82	0.82	0.82	0.86	0.84	0.77	0.85	0.81
	Network action	0.93	0.92	0.92	0.95	0.92	0.94	0.96	0.92	0.94	0.90	0.90	0.90
	Target equipment	0.51	0.58	0.54	0.64	0.63	0.64	0.62	0.63	0.63	0.76	0.58	0.66
	Timeframe	0.75	0.80	0.78	0.81	0.85	0.83	0.85	0.84	0.84	0.98	0.75	0.85
40000	User	0.71	0.72	0.71	0.72	0.70	0.71	0.74	0.69	0.71	0.90	0.94	0.92
	Application utility/goal	0.78	0.86	0.82	0.84	0.87	0.85	0.85	0.85	0.85	0.80	0.85	0.83
	Network action	0.93	0.94	0.93	0.95	0.94	0.94	0.96	0.93	0.95	0.95	0.92	0.94
	Target equipment	0.54	0.55	0.55	0.69	0.65	0.67	0.69	0.65	0.67	0.71	0.66	0.68
	Timeframe	0.83	0.78	0.80	0.86	0.84	0.85	0.87	0.83	0.85	0.99	0.81	0.89

10-30 epochs). Figure 4 considers the results with 1 epoch for BERT and 10 epochs for the benchmarks. As shown in Table 3 and Figure 4a, for all different dataset sizes, the BERT model provides uniformly the best accuracy results as compared to the three LSTM-based schemes (i.e., BiLSTM-CRF, BiLSTM, and LSTM). For instance, with only 1 epoch and datasets of 5000 and 40000 intents, BERT provides the highest accuracy with 80% and 86%, respectively. Meanwhile, with 10 epochs, BiLSTM-CRF, BiLSTM, and LSTM provide 57%, 31%, and 26% for 5000 intents, and 81%, 81%, and 76% for 40000 intents, respectively.

The use of Healthcare 4.0 as a specific case study offers accurate predictions with BERT, even if it is fine-tuned with a small dataset. We can see that BERT significantly outperforms other methods in small datasets (5000 intents), which is a merit since specific data (e.g., Healthcare 4.0) are usually limited, sensitive, and not publicly available for privacy reasons. However, the three LSTM-based techniques are very susceptible to the dataset size, where the higher accuracy results are obtained with the largest dataset (40000 entries). Furthermore, the smaller the dataset size, the more epochs for training are required with these benchmarks. For example, to infer 5000 intents, the LSTM model provides simply 26% of accuracy with 10 epochs and requires 30 epochs to provide 67%. Meanwhile, the fine-tuned BERT model offers 80% accuracy with 5000 intents and only 1 epoch. Thus, the BERT model is independent of the dataset size and starts to converge to the optimal accuracy results from only one epoch. For instance, with 1 epoch and 40000 intents, BERT registers an enhancement of 7%, 5%, and 6% precision, recall, and F1-score, respectively, compared to 5000 intents. In the meantime, the three LSTM-based models' results change intensely from one dataset size to another and from a number of epochs to another.

On the other hand, as expected and displayed in Table 3 and Figure 4b, the processing time for all the implemented schemes augments with the increase of the dataset size. However, the fine-tuning time of the BERT method is longer than the training time of the BiLSTM-CRF, BiLSTM, and LSTM methods due to the high computational complexity. BERT is a large language model including a lot of weights to be updated for each different NLP task. Accordingly, considering the factor of processing time, we fine-tuned BERT with the smallest number of epochs, namely, only one. Generally speaking, the downside of longer processing time, particularly the fine-tuning time, for BERT is not worrisome since this phase can take place only once, or be periodically updated in an offline fashion with new datasets on the Cloud while using multi-core GPUs with parallel systems. From this perspective, it should also be taken into consideration that the generic BERT model is frequently updated by Google, and it incorporates the changes of our daily language. Thus, an occasional fine-tuning of BERT with new healthcare intents will also familiarize the NER model with the neologisms and jargon of healthcare networks, patients, doctors, etc.

Table 4 evaluates how each algorithm performed with each entity of the 5-tuple policy model, presented in Section IV. To do so, the evaluation is performed according to precision, recall, and F1-score metrics for each entity with different dataset sizes and with 1 epoch for BERT and 10 epochs for the benchmarks. It should be noted that similar results were extracted for 20 and 30 epochs, especially as the dataset size increased. However, for illustration purposes and length limitations, only the results with 10 epochs are presented.

Similarly to Table 3, the BERT performance is found to be substantially better as compared to the benchmark models. Specifically, with 5000 intents, BERT provides precision, recall, and F1-score results in the range of [60% - 100%] for all 5 distinct entities with only 1 epoch. The BiLSTM-CRF is the second-best algorithm in terms of classification correctness and provides results equal to or greater than 33%, 20%, and 25% of precision, recall, and F1-score, respectively. These BiLSTM-CRF results are slightly better than BiLSTM (29%, 9%, 15%); meanwhile, LSTM achieves the worst accuracy performance. It is highlighted that BERT has these outcomes because it is based on a general pre-trained model that is later fine-tuned with healthcare network data. It should also be noted that with all dataset sizes, BERT achieves better and more equitable identification results for all entities.

In contrast, the LSTM-based techniques provide unbalance accuracy results for entity identification. For example, the ‘‘target network equipment’’, one of the critical entities in the policy model, registers extremely low (or even zero) accuracy results (represented by dark yellow cells with LSTM, BiLSTM, and BiLSTM-CRF in Table 4). Similar observations could be drawn for the entity ‘‘network action’’ (i.e., QoS, data security, etc.), which provides better performance with the increase of the dataset size compared to the rest of the entities, e.g., from 41%, 78%, 88% of precision with LSTM, BiLSTM, and BiLSTM-CRF, respectively, with 5000 to 95% for the three models with 40000 intents. Meanwhile, the BERT-based model converges to optimal solutions with all entities from 5000 intents to 40000 intents, highlighting its independence from the dataset size for fine-tuning due to its general pre-training characteristic.

In conclusion, processing intents' sequences with a bidirectional approach (from left to right and right to left simultaneously) always outperforms the unidirectional approach (LSTM model). Then, adding a CRF layer over the Bi-LSTM model improves performance since CRF conducts context correlation through a probabilistic graph model while predicting entities. Meanwhile, BERT, the deep bidirectional model, outperforms the other methods and provides the best performance in most of the entities and data sizes due to different merits: Firstly, it is based on a pre-trained model that has accumulated knowledge of huge language corpora. Secondly, it is capable of being fine-tuned on specific domain datasets. Thirdly, it implements the masked language model, which performs word prediction originally hidden in a sequence. Finally, a more technical reason is that BERT leverages the transformer architecture with additional self-attention layers that process the intent's sequence as a whole rather than a word-by-word, uses positional embedding, and estimates the similarity scores of words, even if they are distant in the intent.

VI. CONCLUSION

IBN has become a strong candidate for the automation of current and future network deployments. In this paper, a new IBN-enabled healthcare architecture was introduced, showing the main components and their interaction towards creating an intelligent and flexible smart network. Following, the emphasis was placed on the interfacing of the healthcare network with its users by allowing them to express their network requirements through a human natural language. To this end, and to allow the automatic translation and mapping of the users' utterances to network policies, we designed an intelligent intent refinement system based on the BERT model. The particular model was fine-tuned on a large corpus of labeled entities within 40000 intents specific to smart healthcare network tasks. The attained results revealed that although the processing time of the BERT-based intent refinement method is higher due to the high computational complexity, it considerably outperforms well-known benchmarks in terms of analytical accuracy with only 1 epoch and for a various range of dataset sizes.

It is noteworthy that even though we presented a Healthcare 4.0 use case in this work due to its strong global good related to the quality of human life and well-being, the proposed IBN platform could be deployed over different smart environments. As part of our future work, we plan to implement optimization strategies, such as parallelism and pruning techniques, to enhance the speed of the BERT-based scheme without compromising its accuracy. Then, we intend to analyze the proposed scheme in a real-life testbed (e.g., smart factory and hospital 4.0). Furthermore, we aim to extend our system design by addressing policy conflict resolution and ensuring related services assurance through scheduling and planning schemes. Subsequently, the whole platform will establish the IBN-based closed-loop system, consisting of the three main deployment steps: Refinement, Activation, and Assurance.

ACKNOWLEDGMENT

The authors would like to thank Institut de Cardiologie de Montréal (ICM) and the Centre Hospitalier de l'Université de Montréal (CHUM) for their invaluable input.

REFERENCES

[1] ScienceDaily. *Global Spending on Health is Expected to Increase to $18.28 Trillion Worldwide by 2040*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.sciencedaily.com/releases/2016/04/160414095539.htm

[2] J. Al-Jaroodi, N. Mohamed, and E. Abukhousa, "Health 4.0: On the way to realizing the healthcare of the future," *IEEE Access*, vol. 8, pp. 211189–211210, 2020.

[3] E. Coronado, R. Behravesh, T. Subramanya, A. Fernández-Fernández, M. S. Siddiqui, X. Costa-Pérez, and R. Riggio, "Zero touch management: A survey of network automation solutions for 5G and 6G networks," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2535–2578, 4th Quart., 2022.

[4] Cisco. *Network automation*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.cisco.com/c/en/us/solutions/automation/network-automation.html

[5] IBM. *IBM Knowledge Center*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.ibm.com/us-en?lnk=m

[6] Juniper. *Network Automation*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.juniper.net/us/en/solutions/automation.html

[7] ETSI. *Zero Touch Network & Service Management (ZSM)*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.etsi.org/technologies/zero-touch-network-service-management

[8] M. Behringer, M. Pritikin, S. Bjarnason, A. Clemm, B. Carpenter, S. Jiang, and L. Ciavaglia, "Autonomic networking: Definitions and design goals," Internet Res. Task Force (IRTF), Tech. Rep. RFC: 7575, 2015.

[9] A. Leivadeas and M. Falkner, "A survey on intent-based networking," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 625–655, 1st Quart., 2023.

[10] A. Clemm, L. Ciavaglia, L. Granville, and J. Tantsura, "Intent-based networking-concepts and definitions," Internet Res. Task Force (IRTF), Tech. Rep. RFC: 9315, 2020.

[11] Cisco. *Intent-Based Networking*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.cisco.com/c/en_ca/solutions/intent-based-networking.html

[12] Huawei. *Intent-Based Nemo Overview*. Accessed: Jul. 20, 2023. [Online]. Available: http://www.watersprings.org/pub/id/draft-hares-ibnemo-overview-01.html

[13] Y. Ouyang, C. Yang, Y. Song, X. Mi, and M. Guizani, "A brief survey and implementation on refinement for intent-driven networking," *IEEE Netw.*, vol. 35, no. 6, pp. 75–83, Nov. 2021.

[14] A. Singh, G. S. Aujla, and R. S. Bali, "Intent-based network for data dissemination in software-defined vehicular edge computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 5310–5318, Aug. 2021.

[15] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, and J. Li, "A survey of knowledge enhanced pre-trained language models," *IEEE Trans. Knowl. Data Eng.*, early access, Aug. 30, 2023, doi: 10.1109/TKDE.2023.3310002.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[17] R. A. Addad, D. L. C. Dutra, M. Bagaa, T. Taleb, H. Flinck, and M. Namane, "Benchmarking the ONOS intent interfaces to ease 5G service management," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[18] D. Sanvito, D. Moro, M. Gullì, I. Filippini, A. Capone, and A. Campanella, "Enabling external routing logic in ONOS with intent monitor and reroute service," in *Proc. 4th IEEE Conf. Netw. Softwarization Workshops (NetSoft)*, Jun. 2018, pp. 332–334.

[19] K. Abbas, T. A. Khan, M. Afaq, and W.-C. Song, "Network slice lifecycle management for 5G mobile networks: An intent-based networking approach," *IEEE Access*, vol. 9, pp. 80128–80146, 2021.

[20] X. Zheng, A. Leivadeas, and M. Falkner, "Intent based networking management with conflict detection and policy resolution in an enterprise network," *Comput. Netw.*, vol. 219, Jan. 2022, Art. no. 109457.

[21] A. Yichiet, J. K. Y. Min, G. M. Lee, and L. J. Sheng, "Intent-based network policy to solution architecting recommendations," *Int. J. Bus. Data Commun. Netw.*, vol. 17, no. 1, pp. 55–74, Jan. 2021.

[22] M. Bensalem, J. Dizdarević, and A. Jukan, "Benchmarking various ML solutions in complex intent-based network management systems," 2021, *arXiv:2111.07724*.

[23] M. Bensalem, J. Dizdarevic, F. Carpio, and A. Jukan, "The role of intent-based networking in ICT supply chains," in *Proc. IEEE 22nd Int. Conf. High Perform. Switching Routing (HPSR)*, Jun. 2021, pp. 1–6.

[24] T. Ahmed Khan, K. Abbas, J. J. Diaz Rivera, A. Muhammad, and W.-C. Song, "Applying RouteNet and LSTM to achieve network automation: An intent-based networking approach," in *Proc. 22nd Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Sep. 2021, pp. 254–257.

[25] W. Cerroni, C. Buratti, S. Cerboni, G. Davoli, C. Contoli, F. Foresta, F. Callegati, and R. Verdone, "Intent-based management and orchestration of heterogeneous openflow/IoT SDN domains," in *Proc. IEEE Conf. Netw. Softwarization (NetSoft)*, Jul. 2017, pp. 1–9.

[26] F. Aklamanu, S. Randriamasy, and E. Renault, "Demo: Intent-based 5G IoT application network slice deployment," in *Proc. 10th Int. Conf. Netw. Future (NoF)*, Oct. 2019, pp. 141–143.

[27] S. Badotra, D. Nagpal, S. N. Panda, S. Tanwar, and S. Bajaj, "IoT-enabled healthcare network with SDN," in *Proc. 8th Int. Conf. Rel., INFOCOM Technol. Optim.*, Jun. 2020, pp. 38–42.

[28] Md. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Internet of Things: Device capabilities, architectures, protocols, and smart applications in healthcare domain," *IEEE Internet Things J.*, vol. 10, no. 4, pp. 3611–3641, Feb. 2023.

[29] A. I. Siam, M. A. El-Affendi, A. A. Elazm, G. M. El-Banby, N. A. El-Bahnasawy, F. E. A. El-Samie, and A. A. A. El-Latif, "Portable and real-time IoT-based healthcare monitoring system for daily medical applications," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1629–1641, Jan. 2023.

[30] C. Andrikos, G. Rassias, P. Tsanakas, and I. Maglogiannis, "An enhanced device-transparent real-time teleconsultation environment for radiologists," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 374–386, Jan. 2019.

[31] K. Riad, R. Hamza, and H. Yan, "Sensitive and energetic IoT access control for managing cloud electronic health records," *IEEE Access*, vol. 7, pp. 86384–86393, 2019.

[32] K. Peng, P. Liu, M. Bilal, X. Xu, and E. Prezioso, "Mobility and privacy-aware offloading of AR applications for healthcare cyber-physical systems in edge computing," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 5, pp. 2662–2673, Oct. 2023.

[33] L. Zhang, Y. Zhu, W. Ren, Y. Zhang, and K. R. Choo, "Privacy-preserving fast three-factor authentication and key agreement for IoT-based E-health systems," *IEEE Trans. Services Comput.*, vol. 16, no. 2, pp. 1324–1333, Mar. 2023.

[34] H. Hong, D. Chen, and Z. Sun, "A practical application of CP-ABE for mobile PHR system: A study on the user accountability," *SpringerPlus*, vol. 5, no. 1, pp. 1–8, Dec. 2016.

[35] G. S. Aujla, R. Chaudhary, K. Kaur, S. Garg, N. Kumar, and R. Ranjan, "SAFE: SDN-assisted framework for edge–cloud interplay in secure healthcare ecosystem," *IEEE Trans. Ind. Informat.*, vol. 15, no. 1, pp. 469–480, Jan. 2019.

[36] D. Bringhenti, J. Yusupov, A. M. Zarca, F. Valenza, R. Sisto, J. B. Bernabe, and A. Skarmeta, "Automatic, verifiable and optimized policy-based security enforcement for SDN-aware IoT networks," *Comput. Netw.*, vol. 213, Aug. 2022, Art. no. 109123.

[37] S. N. Matheu, A. Robles Enciso, A. Molina Zarca, D. Garcia-Carrillo, J. L. Hernández-Ramos, J. Bernal Bernabe, and A. F. Skarmeta, "Security architecture for defining and enforcing security profiles in DLT/SDN-based IoT systems," *Sensors*, vol. 20, no. 7, p. 1882, Mar. 2020.

[38] F. Naeem, M. Tariq, and H. V. Poor, "SDN-enabled energy-efficient routing optimization framework for industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5660–5667, Aug. 2021.

[39] G. Aceto, V. Persico, and A. Pescapé, "Industry 4.0 and health: Internet of Things, big data, and cloud computing for healthcare 4.0," *J. Ind. Inf. Integr.*, vol. 18, Jun. 2020, Art. no. 100129.

[40] S. Paul, M. Riffat, A. Yasir, M. N. Mahim, B. Y. Sharnali, I. T. Naheen, A. Rahman, and A. Kulkarni, "Industry 4.0 applications for medical/healthcare services," *J. Sensor Actuator Netw.*, vol. 10, no. 3, p. 43, Jun. 2021.

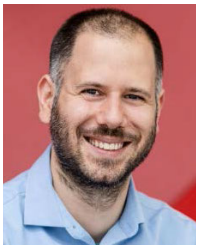
[41] Linux. *Pytorch-Transformers Library*. Accessed: Jul. 20, 2023. [Online]. Available: https://pytorch.org/hub/huggingface_pytorch-transformers/

[42] D. Cournapeau. *Scikit-Learn Library*. Accessed: Jul. 20, 2023. [Online]. Available: https://scikit-learn.org/stable/

[43] Google-Brain. *Tensorflow Library*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.tensorflow.org/

[44] F. Chollet. *Keras Library*. Accessed: Jul. 20, 2023. [Online]. Available: https://keras.io/

YOSRA NJAH (Member, IEEE) received the B.Sc. and M.Sc. degrees in telecommunications from the National Engineering School of Tunis (ENIT), Tunisia, in 2012 and 2013, respectively, and the Ph.D. degree in systems engineering from École de Technologie Supérieure (ÉTS), Montreal, Canada, in 2021. She performed the engineering and master's degrees research projects at the School of Engineering and Architecture of Fribourg (HEIA-FR), Switzerland. She is currently a Postdoctoral Fellow with the Software and Information Technology Engineering Department, ÉTS. Her current research interests include intent-based network automation, software-defined networking, performance optimization, traffic engineering, and network data analytics, with a particular focus on the Internet of Things and data center networks.



ARIS LEIVADEAS (Senior Member, IEEE) received the Diploma degree in electrical and computer engineering from the University of Patras, in 2008, the M.Sc. degree in engineering from King's College London, in 2009, and the Ph.D. degree in electrical and computer engineering from the National Technical University of Athens, in 2015. From 2015 to 2018, he was a Postdoctoral Fellow with the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada. In parallel, he was an Intern with Ericsson and collaborated with Cisco, Ottawa. He is currently an Associate Professor with the Department of Software and Information Technology Engineering, École de Technologie Supérieure (ÉTS), Montreal, Canada. His current research interests include network function virtualization, intent-based networking, cloud and edge computing, the IoT, and network optimization and management. He received the Best Paper Award from ACM ICPE 2018 and 2023 and IEEE iThings 2021; and the Best Presentation Award from IEEE HPSR 2020.

JOHN VIOLOS is currently a Research Associate with the Department of Software Engineering and Information Technology, École de Technologie Supérieure (ÉTS). Previously, he was a Research Associate with the National Technical University of Athens, a Sessional Lecturer with the Harokopio University of Athens, and a Visiting Lecturer with the National and Kapodistrian University of Athens. His current research interests include deep learning, machine learning, and cloud and edge computing. He was a member of the European Commission's Digital Single Market Working Group on the code of conduct for switching and porting data between cloud service providers.

MATTHIAS FALKNER received the M.Sc. degree in operations research and information systems from the London School of Economics and Political Science, U.K., and the Ph.D. degree in systems and computer engineering from Carleton University, Canada. He is currently a Distinguished TME with Cisco's SP Sales CTO Group, where he focuses on enterprise and SP network architectures, particularly on the adoption of 5G technologies for private networks. He also served as a Consulting Systems Engineer for the Deutsche Telekom Account Team. Throughout his career at Cisco, he worked on the evolution of Cisco's Digital Network Architecture (Cisco DNA), the midrange router portfolio, specializing in NFV with the CSR 1000v, the ASR 1000, and IOS XE. His current research interests include intent-based networking, 5G, virtualization, and traffic modeling.