

# Interpretability of the Intent Detection Problem: A New Approach

Eduardo Sanchez-Karhunen <sup>a,\*</sup>, Jose F. Quesada-Moreno <sup>a</sup> and Miguel A. Gutiérrez-Naranjo <sup>a</sup>

<sup>a</sup> Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41012, Sevilla, Andalucía, Spain.

E-mails: fesanchez@us.es, jquesada@us.es, magutier@us.es

**Abstract.** Intent detection, a fundamental text classification task, aims to identify and label the semantics of user queries, playing a vital role in numerous business applications. Despite the dominance of deep learning techniques in this field, the internal mechanisms enabling Recurrent Neural Networks (RNNs) to solve intent detection tasks are poorly understood. In this work, we apply dynamical systems theory to analyze how RNN architectures address this problem, using both the balanced SNIPS and the imbalanced ATIS datasets. By interpreting sentences as trajectories in the hidden state space, we first show that on the balanced SNIPS dataset, the network learns an ideal solution: the state space, constrained to a low-dimensional manifold, is partitioned into distinct clusters corresponding to each intent. The application of this framework to the imbalanced ATIS dataset then reveals how this ideal geometric solution is distorted by class imbalance, causing the clusters for low-frequency intents to degrade. Our framework decouples geometric separation from readout alignment, providing a novel, mechanistic explanation for real world performance disparities. These findings provide new insights into RNN dynamics, offering a geometric interpretation of how dataset properties directly shape a network’s computational solution.

Keywords: Natural Language Processing, Intent Detection, Dynamical Systems, Interpretability

## 1. Introduction

### 1.1. RNNs and the Interpretability Challenge<sup>1</sup>

Modern recurrent neural networks (RNNs) are widely used to tackle problems involving sequential data. These networks have demonstrated strong performance in various natural language processing (NLP) tasks, such as sentiment analysis [34], intent detection and slot filling [16], and machine translation [56]. However, despite their widespread success, the exact nature of the internal mechanisms by which RNNs solve specific tasks remains an open question. This lack of understanding is partly due to the nonlinear nature of RNNs and the high-dimensionality of their hidden layers, which together obscure the computational processes underlying their behavior. Moreover, the trend in practical applications leans towards increasingly complex architectures [4, 56], making it even more challenging to understand what is happening inside the nets. The integration of RNNs into applications with significant societal impact, such as healthcare, legal systems, and autonomous decision-making, has made the demand for interpretability more pressing than ever. Understanding how these models detect patterns, solve problems, and make decisions is no longer merely a theoretical concern but a practical necessity. Enhancing the interpretability of neural networks decision-making is crucial to ensure their robustness, fairness, and accountability in real-world applications [17]. Addressing this challenge is key to bridge the gap between the impressive capabilities of RNNs and the trust required for their deployment in high-stakes environments.

---

\*Corresponding author. E-mail: fesanchez@us.es.

<sup>1</sup>This paper is an extended version of the work presented at 27th European Conference on Artificial Intelligence [48]

## 1.2. Analysis Approaches for Understanding RNNs

Understanding the behavior of RNNs has been a persistent challenge in machine learning research. Early studies focused on visualizing the activity of specific network components, such as memory gates, during NLP tasks [27, 53]. Although these unit-level analyses offer localized functions, they often fail to provide a comprehensive interpretation of the network’s overall behavior. The inherent feedback connections between RNN neurons allow these networks to be viewed as nonlinear dynamical systems [40], allowing the application of well-established tools from dynamical systems theory [54]. Based on this perspective, several studies have derived analytic expressions for aspects of network dynamics such as bifurcations in the parameter space of small networks and convergence properties [20, 60]. These efforts have enhanced our understanding of the mathematical underpinnings of RNN behavior but remain limited in their ability to explain task-specific computations in large, real-world networks. Recently, a new reverse engineering paradigm has emerged to analyze RNNs at a higher level of abstraction. Instead of focusing on the microdetails of individual neurons or gates, this approach examines the state space of trained RNNs. Fixed points are identified and the dynamics of the system is linearized around them, revealing a key computational mechanism embedded within the network [55]. This perspective has yielded significant insights, particularly in the context of text classification tasks, where state space analyses have demonstrated promising results [3, 39]. A recurring theme in these works is the discovery that trained RNNs often converge to highly interpretable, low-dimensional representations associated with attractors in the state space. The geometry and dimensionality of these attractors manifolds are intricately linked to the structure of the dataset and the nature of the task being solved. In summary, this attractor-based view provides a powerful framework for understanding how RNNs encode information and implement computations.

## 1.3. A Comparison with Other Interpretability Paradigms

Our analysis adopts a dynamical systems framework to seek a mechanistic understanding of RNN computation, offering insights distinct from other valuable interpretability paradigms. Contrasting our global, dynamic approach with other common methods clarifies its unique contributions. **Input-attribution methods** aim to attribute a network’s output back to specific input tokens. Attention mechanisms [4], assign relevance scores to input tokens, highlighting what parts of an input sequence the model focuses on when generating an output. Similarly, gradient-based methods like saliency maps [50] compute the output’s sensitivity to small changes in the input, identifying influential features. While these methods excel at answering which input features are most influential, they largely treat the recurrent core as a black box. They offer limited insight into the internal computational processes that transform information into a final classification. **Local surrogate methods:** techniques like LIME [46] and SHAP [38] operate on a different principle. They explain an individual prediction by creating a simpler, interpretable surrogate model that is faithful to the complex model’s behavior in a localized region around the specific input. Their power lies in this instance-specific explanation, but their scope is inherently local and cannot be used to describe the global structure of the state space or the general principles that govern the network’s behavior across all possible inputs.

In contrast, our dynamical systems approach provides a different and complementary level of insight.

- **Revealing dynamic processes:** Unlike static attribution maps, this framework visualizes computation as a dynamic process. It allows us to model sentences as trajectories within the hidden state space, revealing how evidence is accumulated token-by-token, rather than just identifying which tokens were most influential.
- **Characterizing global geometry:** Where other methods are local, our approach characterizes the global geometry of the task learned by the network. This enables the analysis of holistic properties such as the intrinsic dimensionality of the problem and the organization of the decision space into meaningful regions.
- **Uncovering mechanistic principles:** Most critically, this perspective moves from correlation to mechanism. It seeks to explain why the network functions as it does, framing classification as a process governed by the stable and unstable dynamics of a fixed-point topology. Therefore, our approach complements other methods by providing a unique lens into the internal, dynamic, and geometric nature of how RNNs solve tasks.

#### *1.4. Intent Detection: A Case Study Domain*

Within the field of NLP, intent detection is a horizontal foundational operation, providing instrumental support for a wide range of applications. Broadly defined, intent detection is an NLP task aimed at recognizing and classifying the underlying purpose or operational goal (a.k.a. intention) expressed in a user’s utterance. It serves as a critical component in the functional architecture of language understanding systems [43], addressing the challenge of mapping a large and diverse set of linguistic expressions onto a predefined set of semantic intentions. This challenge is complex, requiring the operation of multiple linguistic levels simultaneously. From diverse lexical realizations and syntactic structures, to semantic ambiguities and pragmatic contexts, intent detection must integrate diverse dimensions of language processing to deliver accurate results. Despite progress in academic research and industrial applications, intent detection remains a theoretical and practical challenge. Different lines of interest include: detecting multiple intents within a single utterance [30], integrating intent detection with entity recognition [2, 51, 58], managing out-of-domain intents [7, 32], and developing robust models that support explainability and transparency in classification decisions [62]. These challenges, combined with the intrinsic characteristics of intent detection: a) operation on a low-dimensional semantic space, despite the high-dimensionality of the architectures involved, and b) a topologically inspired convergence towards different semantic kernels, make it an ideal domain for analysis using dynamical systems theory. This work leverages a combined mathematical, computational, and linguistic framework rooted in dynamical systems theory to propose a novel approach to understanding the nature of the intent detection process. This framework also constitutes the starting point for tackling some of the challenges indicated.

#### *1.5. Our Contributions and Paper Organization*

Our key contribution is the pioneering study of the state-space dynamics of trained RNNs applied to the SNIPS and ATIS intent detection problems. We show that the state space can be characterized as a low-dimensional manifold whose intrinsic dimensionality is related to the size of the embedding layer and the number of neurons in the hidden layer. We show that input sentences traverse discrete trajectories through the state space, progressing from initial states toward specific outer regions. A crucial finding is the identification of distant regions within the state space, where these trajectories terminate. These peripheral areas are aligned with the directions defined by the rows of the readout matrix, allowing for the generation of predictions. In addition, we uncover the fixed-point topology underlying the network dynamics. Unlike other tasks, such as sentiment analysis or document classification, we find that RNNs trained for intent detection exhibit an unexpected fixed-point structure [3]. The number and nature of attractors, saddle points and other critical points in the state space are shown to depend on network parameters and the type of RNN cell (e.g. LSTM or GRU). We first established this geometric framework on the balanced SNIPS dataset, and then use it as a diagnostic tool to test its generalizability on the complex, imbalanced ATIS dataset. This analysis reveals how the ideal geometric solution is distorted by class imbalance on low-frequency intents. We introduce a novel diagnostic framework that decouples geometric separation from readout alignment, allowing us to identify four distinct, mechanistic patterns that explain real-world performance disparities.

The rest of the paper is organized as follows: Section 2 introduces the problem of intent detection, discussing various aspects and current lines of research in this field, as well as how the dynamical systems approach fits into this context. Section 3 explores how RNNs can be interpreted as nonlinear dynamical systems. Section 4 outlines the specific objectives to be addressed through the experiments. Section 5 describes the datasets selected for this study. Section 6 details the experimental setup and methodology, while Section 7 provides an in-depth analysis of the results obtained. Finally, Section 8 presents the conclusions and suggest potential directions for future research.

## **2. Intent Detection: Domain Characterization and Research Challenges**

Intent detection is a cornerstone in several areas of NLP, playing a vital role in numerous industrial applications, particularly in conversational systems, virtual assistants, and question-answering environments. At its core, intent detection addresses one of the most fundamental syntactic-semantic challenges in natural language understanding: to collaborate effectively in conversational interactions, a conversational agent must accurately identify and unambiguously classify the user’s intent. This task depends on various linguistic components and levels. From lexical and morphological units to syntactic, semantic, and even pragmatic and acoustic nuances. In particular, even non-verbal cues, such as silence or pauses, can convey relevant intentional meanings.