IES International Journal of Multidisciplinary Engineering Research

Foundations and Methodologies of Rule-Based Semantic Role Labelling: A Review

JIJI A J

Assistant Professor, Department of Computer Science & Engineering(Data Science), IES College of Engineering, Kerala, India

Email_id: jiji@iesce.info

Abstract

Semantic Role Labeling (SRL) provides a structured semantic interpretation of natural language by identifying predicate–argument relationships that encode event-level information such as agentivity, affected entities, location, and temporality. As a foundational layer in the NLP pipeline, SRL leverages syntactic features from part-of-speech tagging, chunking, and dependency parsing to generate semantic representations that support downstream tasks including question answering, information extraction, and machine translation. This review presents a detailed technical analysis of traditional rule-based SRL, which operationalizes linguistic theory through deterministic mappings between syntactic configurations and semantic roles. The methodology encompasses sentence segmentation, tokenization, POS tagging, lemmatization, predicate detection, syntactic dependency analysis, rule application, and role validation. Semantic roles are assigned through handcrafted patterns derived from grammatical relations such as nsubj, obj, iobj, and specific prepositional structures. While rule-based SRL ensures interpretability, structural consistency, and independence from annotated corpora, its reliance on manually engineered rules restricts generalization to diverse or ambiguous linguistic inputs. This paper rigorously evaluates these mechanisms, highlighting the architectural characteristics, strengths, and inherent limitations that define rule-based SRL within modern computational linguistics.

Keywords: POS Tagging, Dependency Parser.

DOI: https://doi.org/10.5281/zenodo.19232769

1. Introduction

Semantic Role Labeling (SRL) is fundamentally connected to the broader field of Natural Language Processing (NLP), as it provides the semantic layer of understanding required for many higher-level language technologies. NLP aims to enable machines to interpret, generate, and reason about human language, and SRL contributes directly by identifying the predicate–argument structures that encode who did what, to whom, when, where, and how within a sentence. This mapping from surface text to semantic roles supports the transition from syntactic processing to deeper semantic interpretation. SRL serves as a bridge between structural linguistic analysis and meaning representation. NLP tasks such as part-of-speech tagging, chunking, and syntactic parsing provide essential input features—such as phrase boundaries, dependency relations, and verb frames—that SRL relies on to detect predicates and their associated arguments. In turn, the semantic representations produced by SRL enhance downstream NLP applications, including information extraction, question answering, machine translation, text summarization, and sentiment analysis. By transforming unstructured text into structured semantic information, SRL enables NLP systems to perform more

accurate reasoning and decision-making.

Early Semantic Role Labeling systems were predominantly rule-based, relying on handcrafted linguistic rules, verb subcategorization frames, and syntactic heuristics to identify predicate–argument structures. These models were interpretable and required no annotated datasets, but they were difficult to scale, highly domain-dependent, and unable to generalize to complex or unseen sentence patterns. In a typical rule-based SRL system, linguistic knowledge is operationalized through explicit patterns that map syntactic relations—such as subjects, objects, and prepositional phrases—to semantic roles like Agent, Patient, Location, or Temporal expressions. While these systems perform reliably on well-formed, grammatically predictable text, their rigidity limits their adaptability, especially in the face of linguistic variability or ambiguity.

2. Related Work

Early work in SRL grew from the development of lexical-semantic resources such as FrameNet, PropBank and VerbNet, which provided the theoretical and annotated foundations for mapping predicates to argument structures. FrameNet formalized frame-semantic representations based on frame elements and their realizations in text, while PropBank offered a predicate-centered annotation scheme widely used in SRL benchmarking. These resources not only enabled consistent annotation but also informed the design of rule-based systems that operationalize linguistic generalizations for role assignment.

In Rule-based and knowledge-driven SRL[1][2] before large annotated corpora and robust statistical learners became widespread, many systems relied on handcrafted rules, lexica and subcategorization frames to extract predicate–argument relations. Rule-based systems exploited syntactic cues, lexical frames and deterministic pattern matching to assign roles; they are valued for interpretability and for requiring little or no labeled data. In Statistical and feature-based SRL[2] with the availability of PropBank[15] and shared evaluation tasks (e.g., CoNLL)[8][16], SRL research moved to supervised learning. Early statistical models used rich hand-crafted features derived from POS tags, syntactic parses, and lexical resources and applied classifiers such as maxent and SVMs to detect and classify arguments. Thematic roles, originating from the work of Fillmore (1968) and Gruber (1965), provide a structured way to represent these semantic relationships between verbs and their arguments. While there is no universal set of thematic roles, commonly used roles include AGENT, EXPERIENCER, FORCE, THEME, RESULT, CONTENT, INSTRUMENT, BENEFICIARY, SOURCE**,** and GOAL,[4][6][7] each describing a specific type of participant in an event. Different computational models may use small, abstract sets of semantic roles or larger, more fine-grained sets depending on the level of semantic detail required.

In neural end-to-end SRL[3] the advent of deep learning produced a leap in SRL performance. Neural architectures replaced manual feature engineering with dense, learned representations. Bidirectional LSTMs with CRF[18] or structured decoding layers became a dominant paradigm, producing strong results on span- and dependency-based benchmarks. In Sequence-to-sequence and generative formulations[4] more recently, seq2seq architectures have been explored for SRL, treating annotation as a generation task that produces linearized bracketed role representations. The rule-based methodology reviewed in this paper both represents the historical foundation of SRL and continues to inform contemporary efforts that seek to combine linguistic rigour with the adaptability of data-

driven models.

3. Methodology

Semantic Role Labelling identifies who did what to whom, when, where, and how in a sentence. A rule-based SRL system uses manually created linguistic rules—not machine learning—to assign semantic roles such as Agent, Patient, Instrument, Location, Time[6][7], etc. Rule-based SRL is one of the earliest approaches to semantic parsing. Instead of training on annotated corpora, it relies on grammar rules, syntactic patterns, and lexical resources to identify the predicate and label the arguments. It provides high interpretability and reliable performance for well-structured text but struggles with noisy or ambiguous sentences. The Architecture diagram shown in figure 3.1.

The Steps in rule based semantic role labelling are the following:

1. Text Preprocessing
2. Predicate Identification
3. Syntactic Parsing
4. Rule Application & Pattern Matching
5. Semantic Role Assignment
6. Role Validation & Post-processing

In text Preprocessing, the first essential task in this process is sentence segmentation, which involves identifying individual sentence boundaries within a continuous text. Once the text is divided into individual sentences, the next phase, tokenization, decomposes each sentence into its smallest meaningful units known as tokens. Tokens include words, numbers, punctuation marks, and occasionally symbols depending on the tokenizer design. Tokenization plays a critical role in SRL because nearly all subsequent tasks—parsing, role assignment, and rule matching—operate at the token level. Effective tokenization requires handling complexities such as contractions (“didn’t” → “did” + “n’t”), hyphenated expressions (“state-of-the-art”), and multi-word entities (“New York”). A well-designed tokenizer preserves the linguistic integrity of each sentence while ensuring that every token is isolated as a discrete analytical unit for syntactic processing.

Following tokenization, part-of-speech (POS) tagging assigns a grammatical category to each token, identifying whether a word acts as a noun, verb, adjective, adverb, preposition, or other syntactic role. POS tagging is indispensable in rule-based SRL because the identification of predicates, arguments, and semantic roles depends strongly on grammatical structure. Verbs generally serve as predicates around which semantic frames are constructed, while nouns typically function as potential arguments for roles such as Agent, Patient, or Location. Prepositions also hold significant importance, as they frequently signal semantic roles like temporal, locative, or beneficiary relations. POS tagging systems may operate through handcrafted grammatical rules, statistical models, or hybrid approaches, but their primary objective remains consistent: to provide a syntactic profile for each token that guides the later stages of semantic interpretation. Accurate tagging is essential, as grammatical ambiguity (such as “book” functioning as either a noun or a verb) can significantly impact downstream role assignment. The final major preprocessing task is lemmatization, which converts words into their canonical or dictionary form. Lemmatization is especially vital in SRL because many semantic role rules and predicate definitions rely on base verb forms.

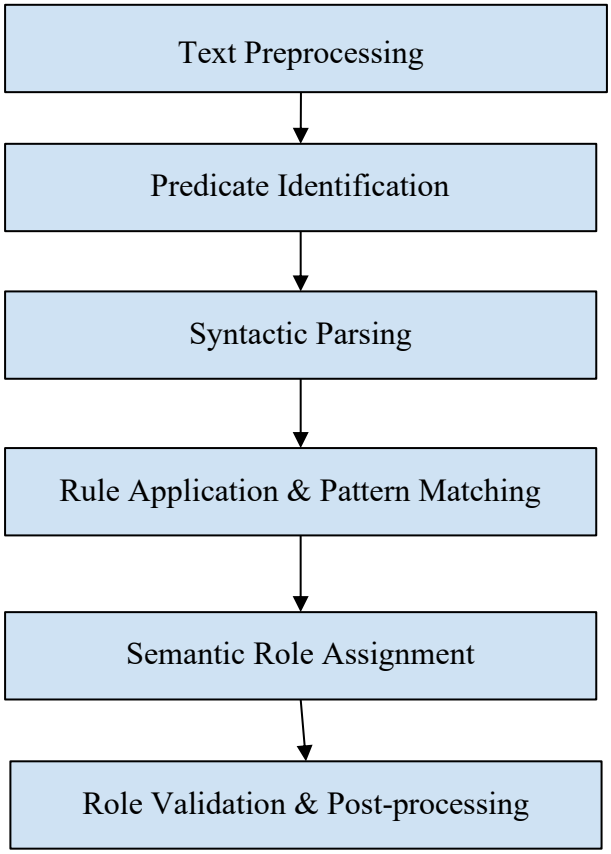

Figure 1: Architecture Diagram

For example, the words “cooking,” “cooked,” and “cooks” all correspond to the lemma “cook,” and only by reducing these variations to their root form can an SRL system accurately match them with verb frames in lexical resources such as VerbNet or FrameNet. Unlike stemming, which simply strips suffixes without considering linguistic correctness, lemmatization incorporates morphological analysis and part-of-speech information to ensure accurate normalization. This semantic standardization of tokens supports consistent predicate identification and argument mapping, even when the surface forms of words vary due to tense, number, or derivation.

Together, sentence segmentation, tokenization, POS tagging, and lemmatization create a structured, linguistically coherent representation of the input text. This representation forms the basis upon which syntactic parsing and semantic role rules can reliably operate. High-quality preprocessing therefore directly influences the accuracy, consistency, and interpretability of any rule-based Semantic Role Labelling system.In Predicate Identification, predicate identification represents one of the most crucial stages in the Semantic Role.

Labelling (SRL) pipeline, as it forms the foundation upon which the entire semantic interpretation of a sentence is constructed. In linguistic theory, the predicate—most commonly realized as the main verb—functions as the central semantic unit that determines the event or action being described. It governs the number and types of arguments that may be involved and provides the structural and semantic frame that role labels such as Agent, Patient, Location, and Instrument attach to. Accurate identification of the predicate is therefore essential, as any error at this stage propagates through subsequent phases of SRL, ultimately compromising the integrity of semantic role

assignment.

In rule-based SRL systems, predicate identification relies on a combination of syntactic, morphological, and lexical cues. The most basic approach involves detecting verbs through part-of-speech tags, particularly tags such as VB, VBD, VBG, VBN, and VBZ, which directly signal verbal forms. However, predicate identification extends beyond merely locating any verb in a sentence; it requires recognizing which verb functions as the main predicate and distinguishing it from auxiliary verbs, modal verbs, and verbal modifiers. For example, in the phrase "has been eating," the true predicate is "eating," while "has" and "been" serve auxiliary grammatical functions supporting tense and aspect. Rule-based systems therefore incorporate heuristics and grammatical patterns to differentiate between main verbs and supporting auxiliaries.

Sentence: "The teacher explained the lesson..."

Tokens → POS Tags → Identify Verb → Mark as Predicate

explained → VBD → Predicate

In Syntactic Parsing, a syntactic parse tree or dependency tree is generated to uncover the structural relations between words. A dependency parser is an NLP system that analyzes the grammatical structure of a sentence by identifying head–dependent relationships between words rather than grouping them into phrase structures. In dependency parsing, each word (except the root) is linked to another word that governs it, forming a directed graph or tree. The resulting structure shows how words depend on one another to convey meaning—for example, identifying subjects, objects, modifiers, and complements.

Unlike constituency parsing, which builds hierarchical phrase chunks (NP, VP, PP), a dependency parser directly connects words through labeled arcs such as nsubj (nominal subject), obj (object), amod (adjective modifier), or det (determiner). This makes dependency structures compact, intuitive, and effective for semantic tasks.

A typical dependency representation for "The dog chased the cat" is shown inFigure 3.2.

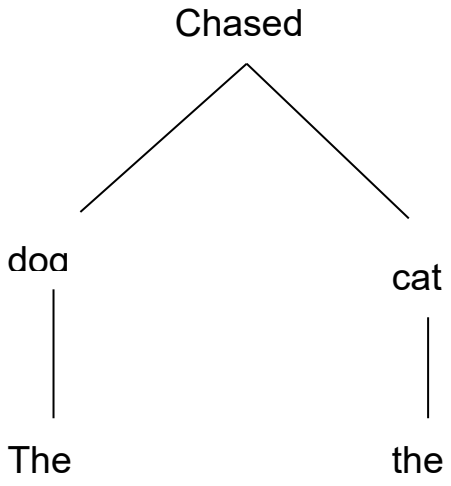

Figure 2: Dependency Graph

Rule Application & Pattern Matching

Handcrafted rules map syntactic relations to semantic roles.[4][6][7]

1. Agent (doer of the action)

Example rule:If the token has relation nsubj(verb, X), mark X as AGENT.

Example:"John baked a cake." → John = AGENT

2. Patient (entity affected by the action)

Example rule:obj(verb, X) ⇒ PATIENT

Example:"John baked a cake." → cake = PATIENT

3. Beneficiary (receiver of the benefit)

Example rule:If a PP with "for/to" modifies the verb ⇒ BENEFICIARY.

Example:"John baked a cake for Mary." → Mary = BENEFICIARY

4. Temporal (time-related information)

Example rule:If PP expresses time ⇒ TEMPORAL.

Example:"He left at 5 pm." → 5 pm = TEMPORAL

5. Location (place where action occurs)

Example rule:If PP expresses spatial location ⇒ LOCATION.

Example:"The kids played in the park." → park = LOCATION

Semantic Role Assignment is the process of determining the underlying meaning-based roles that different constituents in a sentence play with respect to the main predicate (usually a verb). Once the predicate is identified, the system assigns roles such as Agent, Patient, Experiencer, Instrument, Beneficiary, Location, and Temporal based on syntactic structure and semantic patterns.At its core, semantic role assignment answers the question:

"Who is doing what, to whom, when, where, and how?"

In Role Validation & Post-processing ensure that the assigned semantic roles are consistent, non-conflicting, and structurally well-formed before producing the final SRL output. This stage checks for errors such as duplicate core roles, missing required arguments, or incorrect span boundaries, and applies correction rules to refine the semantic interpretation.For the sentence "The chef cooked dinner for the guests in the kitchen," the system identifies the predicate cooked and assigns roles such as Agent (the chef), Patient (dinner), Beneficiary (the guests), and Location (in the kitchen)[4][6][7], illustrating how rule-based SRL maps syntactic elements to meaningful semantic roles.

4. Results

Role	Meaning	Typical Syntactic Pattern	Example
Agent	Doer of the action	Subject (*nsubj*)	***John*** opened the door.
Patient	Entity affected	Object (*obj/dobj*)	John opened ***the door***.
Beneficiary	Receiver of a benefit	"for", "to", *iobj*	John baked a cake ***for Mary***.
Instrument	Tool used	"with" prepositional phrase	He cut it *with a* ***knife***.

Location	Where the event occurs	PP: *in, at, on*	They met ***at school.***
Temporal	When the event occurs	PP: *at, during, before*	He arrived ***at noon.***
Experiencer	Entity that perceives or feels	Subject of mental-state verb	***She*** felt happy.

Table 1: Semantic Roles With Examples

Table 1 represents the semantic roles , meaning, typical synaptic patterns and also examples of each role. The table presents a structured overview of the most commonly used semantic roles in Semantic Role Labeling (SRL), illustrating how each role corresponds to a specific type of participant in an event. For each role, the table outlines its meaning, the typical syntactic pattern that signals the role in a sentence, and a clear example demonstrating its usage.

Roles such as Agent and Patient capture core event participants—the doer of an action and the affected entity—while roles like Beneficiary, Instrument, Location, and Temporal represent contextual details that specify who benefits, what tool is used, where an event occurs, and when it takes place. The Experiencer role highlights entities involved in mental or perceptual processes. Collectively, the table provides a concise reference for mapping syntactic cues to underlying semantic interpretations, forming an essential component of rule-based SRL systems.

5. Conclusion

Rule-based Semantic Role Labeling represents one of the earliest and most linguistically grounded approaches to extracting predicate–argument structures from natural language. By leveraging explicit grammatical rules, syntactic relations, and lexical resources, these systems provide transparent and interpretable mappings from surface syntax to underlying semantic roles. The methodology—from text preprocessing and predicate identification to syntactic parsing, rule application, and final role validation—demonstrates a highly structured and deterministic pipeline capable of producing consistent semantic analyses for well-formed text. However, the reliance on manually engineered rules also imposes limitations, particularly in handling linguistic variability, ambiguous constructions, and domain-specific language patterns.

Despite these challenges, rule-based SRL holds enduring value in NLP, both as a foundation for understanding semantic structure and as a benchmark for evaluating more flexible, data-driven models. Its clarity and linguistic rigor contribute significantly to semantic interpretation, making it an essential precursor to modern neural SRL techniques. Understanding the strengths and limitations of rule-based SRL[1] not only provides historical perspective but also highlights the need for hybrid and learning-based approaches that can combine linguistic precision with the robustness and adaptability required in contemporary language technologies.

6. References

[1]. Rashmi R. Chouhan, Dr. Charmy S Patel 30, 2024. A Review On Semantic Role Labelling For Indian Languages Educational Administration: Theory and Practice, 2148-2403, 15958-15999 https://www.researchgate.net/publication/384747769_A_Review_On_Semantic_Role_Labelling_For_Indian_Languages

[2]. Sunitha C, Dr. A Jaya, Amal Ganesh, 30th April 2018. Semantic Role Labelling of Malayalam Web Documents in Cricket Domain, Journal of Theoretical Information Technology and Applied, vol.96, No.8, 1992-8645.

https://www.jatit.org/volumes/Vol96No8/13Vol96No8.pdf

[3]. Daniel Jurafsky, James H Martin, Semantic Role Labelling. 24 August 2025, Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/21.pdf

[4]. Manfred Klenner, Anne Göhring, Semantic Role Labelling for Sentiment Inference, Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022).

[5]. Daniil Larionov, Artem Shelmanov, Elena Chistova, Ivan Smirnov, sep 2019, Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates, Proceedings of Recent Advances in Natural Language Processing, 619-622 https://aclanthology.org/R19-1073.pdf/

[6]. Mihai Surdeanu, Llu´ıs M`arquez, Xavier Carreras, Pere R. Comas, June 2007, Combination Strategies for Semantic Role Labeling, Journal of Artificial Intelligence Research 29 (2007) 105-151.https://www.jair.org/index.php/jair/article/view/10500/25157 https://www.sciencedirect.com/topics/computer-science/semantic-role-labeling

[7]. Betina Antony J, G. Suryanarayanan Mahalakshmi, July 2015, Content-based Information Retrieval by Named EntityRecognition and Verb Semantic Role Labelling, Journal of Universal Computer Science, Vol. 21, 1830-1848 https://lib.jucs.org/article/23832/

[8]. eorgios Petasis, Frantz Vichot , Francis Wolinski, Georgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos, Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems, Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos". https://www.analyticsvidhya.com/blog/2021/06/rule-based-sentiment-analysis-in-python/ https://www.sciencedirect.com/science/article/pii/S2667305325001255

[9]. Shailja Gupta, Rajesh Ranjan, Surya Narayan Singh, Comprehensive Study on Sentiment Analysis: From Rule based to modern LLM based system. https://medium.com/aimonks/semantic-role-labeling-unveiling-the-meaning-behind-language-a4d48d4986af, https://propbank.github.io/ https://universaldependencies.org/conll18/

[10]. Diego Marcheggiani , Anton Frolov , Ivan Titov, August 3 - August 4, 2017, A Simple and Accurate Syntax-Agnostic Neural Model for Dependency-based Semantic Role Labeling, Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 411–420. https://www.geeksforgeeks.org/machine-learning/understanding-of-lstm-networks/ https://aclanthology.org/K17-1041.pdf

[11]. Luheng He , Kenton Lee , Mike Lewis , and Luke Zettlemoyer, Deep Semantic Role Labeling: What Works and What's Next, July 30 - August 4, 2017, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,, pages 473–483. https://www.geeksforgeeks.org/nlp/semantic-roles-in-nlp/

[12]. Martha Palmer, Daniel Gildea, Paul Kingsbury, 9th December 2003, The Proposition Bank: An Annotated Corpus of Semantic Roles, Association for Computational Linguistics, Volume 31, Number 1.