Innovation Humanities and Social Sciences Research (IHSSR)
Volume 22, Issue 4(2026). ISSN：2949-1282

A Machine Learning Study of Chinese Emphatic "shì" (是) Based on Event Structure and Deep Algorithms

Hongyu Nan[1], Haiying Zhang[2]

[1] School of International Education, Yancheng Teachers University, Yancheng 224000, China
[2] School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114000, China

ABSTRACT
This paper proposes a deep learning-based approach for automatically generating Chinese sentences containing the emphatic "是" (shì), grounded in the relationship between its syntactic behavior and the extraction of semantic roles from event structures. We develop a two-stage neural model: a BERT-based semantic role labeler first identifies key semantic components, which then serve as constraints for a Transformer-based conditional text generator. The proposed method produces grammatically correct emphatic "是" sentences and achieves strong performance in both automatic benchmarks and human evaluations. This study offers a novel framework for computationally modeling Chinese modality.

Keywords: *Emphatic "是", Event structure, Machine learning, Semantic role labeling, Chinese syntax generation.*

1. INTRODUCTION

Syntactic research in linguistics provides a foundation for natural language information processing. Its primary goal is to reveal the inherent expressive rules of language through structural analysis and to formalize these rules for computational processing. This enables computers to automatically derive the grammatical structure of sentences, identify the syntactic units they contain, and determine the relationships among those units.

Traditional Chinese linguistics has primarily regarded "是" as a copula expressing judgment (Wang Li，2011). However, substantial linguistic evidence shows that "是" can also convey rich modal meanings, especially an emphatic function. A growing body of Chinese linguistic research recognizes this function as an important means by which Chinese speakers express subjective attitudes and highlight informational focus (Fang Mei 1995; Zhang Heyou 2007; Zhou Guozheng 2008; Jin Lixin 2013; Wang Ping 2019). Existing research on the emphatic function of "是" has mainly focused on its semantic and pragmatic dimensions, whereas studies of its syntactic generation remain limited, particularly those using machine learning-based approaches.

2. AN EVENT STRUCTURE-BASED SYNTACTIC MODEL OF CHINESE EMPHATIC "是"

This paper builds on the analysis of Hongyu Nan and Liming Yu (2019), which uses event structure theory to explain the emphatic function of "是" in Chinese. Existing academic research on the emphatic function of "是" is largely grounded in the "sentence-centric" hypothesis. This view holds that "是" serves as a contrastive focus marker, highlighting semantic roles such as agent, time, location, and instrument, but generally does not mark patient arguments following the verb (Fang Mei 1995). However, this analysis faces two major challenges:

First, it lacks a systematic, principle-based explanation of the focus-marking capacity of "是". Linear analyses have not fundamentally clarified their licensing conditions or constraints. In fact, beyond agents, time, location, and instrument, "是" can also mark semantic roles such as manner and action, as in the following examples:

- 皇后起先是亲自出面替李处温解释。（方式）（徐兴业《金瓯缺》）

  Huánghòu qǐxiān **shì** qīnzì chūmiàn tì Lǐ Chǔwēn jiěshì. (Manner)

  The Empress initially appeared in person to explain

  Li Chuwen's behalf. (Xingye Xu, The Broken Golden Bowl)

- （我）后来仿佛是来到了水边。（行为）（余华《活着》）

  (Wǒ) hòulái fǎngfú **shì** lái dào le shuǐ biān. (Action)

  Later, it seemed that I had arrived at the water's edge. (Yu Hua, To Live)

Second, the sentence-centric account does not provide a unified explanation of the exhaustive focusing of syntactic constituents. Although the claim that "是" cannot focus post-verbal patients is accurate, the "的是" construction can successfully focus such elements, forming a comprehensive focus-marking system through complementary distribution:

- 楼下的铺面房开的是烟草店。（徐兴业《金瓯缺》）

  Lóuxià de pùmiànfáng kāi **de shì** yāncǎodiàn.

  The shopfront downstairs operated as a tobacco store.

  (Xu Xingye, *The Broken Golden Bowl*)

- （咱们）这次造的是官桥。（余华《在细雨中呼喊》）

  (Zánmen) zhè cì zào **de shì** guānqiáo.

  What we are building this time is an official bridge.

  (Yu Hua, *Cries in the Drizzle*)

This indicates that a purely linear analysis of contrastive focus is insufficient to explain the generative logic underlying the emphatic function of "是". A systematic account of the marking capacity of "是", its constraints, and its complementary distribution with the "的是" construction requires closer examination of its underlying semantic structure and of the interface mechanisms between syntactic generation and semantic representation.

Hongyu Nan and Liming Yu (2019) proposed a sentence-generation theory based on event structure, offering a new perspective for explaining the emphatic function of "是". Their analysis suggests that the underlying foundation of Chinese is a logical event structure composed of semantic roles. The emphatic "是" belongs to the modal layer, positioned between the syntactic structure layer and the event structure layer. Its function is to extract a specific semantic role from the event structure and perform a focalization operation on it within the syntactic structure, as illustrated in "Figure 1".

"Figure 1" illustrates a syntactically minimalist and systematic operational process for the emphatic function of "是" in Chinese. On the one hand, it explains why "是" can flexibly precede various syntactic constituents. On the other hand, it demonstrates the syntactic realization of patient-role extraction through complementary focalization. Based on "Figure 1", the sentence-pattern modeling is shown in "Figure 2".

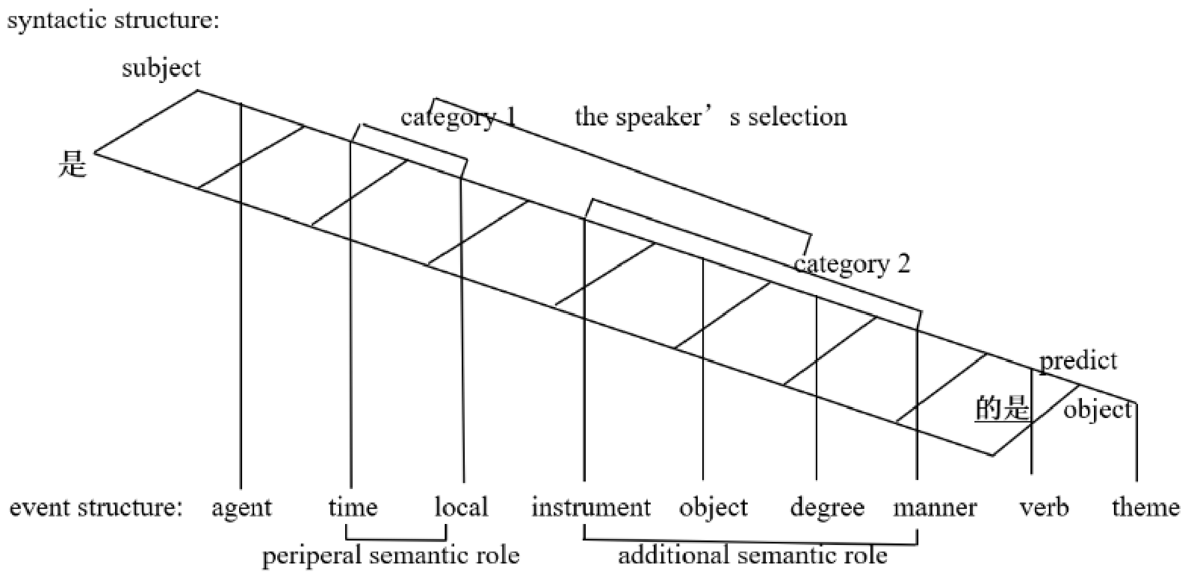

Figure 1 Event semantic extraction mechanism with an emphasis on "shi".

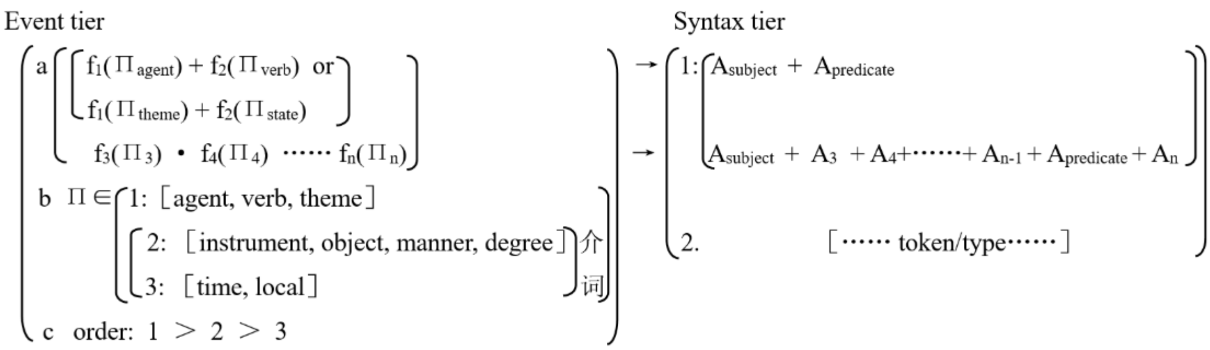

Figure 2 The syntactic generation function of event structure.

"Figure 2" constructs an event-driven framework for syntactic generation. The event layer defines the semantic structure through function composition, hierarchically annotates semantic roles into core, adjunct, and peripheral layers, and specifies their default order. The syntactic layer then maps the event structure onto a linear syntactic sequence through a series of combinatorial operations (e.g., A1...n), ultimately inserting specific lexical items to form the surface sentence. Based on the Chinese sentence-generation mechanism constructed in "Figure 2", and using the symbol "Π" to represent semantic roles, the syntactic generation of the emphatic "是" in Chinese can be expressed as follows:

If [是↠ f(Π)], then [是+ A_n]

(Constraint: If Π ∈ Theme, "是" is projected as "的是".)

After the semantic roles in the event structure are projected into their corresponding positions in the syntactic structure, a syntactic slot exists before each constituent. The emphatic "是" can occupy these syntactic slots. An analysis based on the extraction of event semantic roles by the emphatic "是" enables a unified operation for both patient and non-patient roles. This approach achieves two goals: (1) it satisfies the semantic-category attribution of the focused element, and (2) it provides a unified operation for extracting semantic roles from event structure and projecting them into syntactic structure.

3. A TWO-STAGE DEEP LEARNING MODEL FOR CHINESE EMPHATIC "是"

At the algorithmic level, we define the input as an original Chinese sentence S and the output as a grammatical Chinese sentence S′ containing the emphatic "是", where the emphasized semantic role is Π.

3.1 Stage 1: Semantic Role Labeling (SRL)

This study employs a BERT-based Semantic Role Labeling (SRL) model to extract structured semantic role information from the input sentence. By integrating deep semantic representations with surface syntactic structures, the model improves the semantic fidelity and syntactic naturalness of the generated text, thereby providing robust support for event structure-driven language generation. Specifically, the BERT-based SRL model is used to extract the semantic role set R from the event structure of the input sentence S:

R=BERT-SRL(S)

In this formula, we define:

$R=\{(\Pi_i, \text{type}_i, \text{span}_i)\}$

where:

Π_i: the content of the semantic role

type_i: the type of semantic role (e.g., ARG0, ARG1, ARGM-TMP, ...)

span_i: the position span of the semantic role within the sentence

SRL identifies not only predicates but, more importantly, the corresponding arguments for each predicate, assigning semantic role labels such as ARG0, ARG1, and ARGM-TMP. The semantic role labeling process is described below:

The objective of semantic role labeling is to identify predicates in a sentence on the semantic basis of event structure. The SRL process is as follows:

- Sentence: 现在 我 在一个 小 房间 里 写 这封 信。

  Xiànzài wǒ zài yīgè xiǎo fángjiān lǐ xiě zhèfēng xìn.

  Now I am writing this letter in a small room.

Target Predicate: 写 (write)

Logical Semantic Representation:

- (∃e)[写(e) ∧ ARG0(e, 我) ∧ ARG1(e, 信) ∧ ARGM-TMP(e, 现在) ∧

  ARGM-LOC(e, 小房间里)]

Semantic Annotation Results:

我 = ARG0; 写 = Predicate; 信 = ARG1;

现在 = ARGM-TMP; 小房间里 = ARGM-LOC

Workflow of the BERT-based SRL Model:

Input Encoding: The sentence is tokenized and fed into the BERT model to generate contextualized embeddings for each token.

Predicate Identification: The model first identifies the target predicate "写" (write) as the core of the event.

Argument Labeling: For the identified predicate, the model classifies each token or span into a semantic role:

- "我" (I) is labeled as ARG0 (Agent).

  "信" (letter) is labeled as ARG1 (Theme/Patient).

  "现在" (now) is labeled as ARGM-TMP (Temporal adjunct).

  "小房间里" (in a small room) is labeled as ARGM-LOC (Locative adjunct).

Structured Output: The model outputs the structured semantic roles as a set of tuples:

- R = {(我, ARG0, [0]), (信, ARG1, [5]), (现在, ARGM-TMP, [1]), (小房间里,

  ARGM-LOC, [3])}

The Semantic Role Labeling (SRL) system serves as the initial processing stage by extracting structured semantic role information from input sentences. The detailed workflow of the BERT-based SRL model is presented below.

- a. Predicate Identification and Disambiguation:

We first identify all predicate candidates, primarily verbs, in the sentence. For each candidate, the model performs contextual disambiguation to determine its specific sense, thereby activating the corresponding argument-structure template.

- b. Argument Identification and Role Classification:

For each disambiguated predicate, the model jointly performs argument identification and role classification. We adopt the BIO (Begin-Inside-Outside) tagging scheme to demarcate argument spans. Based on BERT-encoded representations, a classification network predicts a label (e.g., B-ARG0, I-ARG1, or O) for each token in the sequence, thereby precisely determining the boundaries and types of semantic roles.

The accurate semantic role information obtained through this pipeline serves as a strong semantic constraint for the subsequent syntactic generation module, ensuring both semantic fidelity and syntactic naturalness in the generated text.

3.2 Stage 2: Conditional Syntax Generation

We employ a Transformer-based generative model (e.g., T5, BART, or GPT with control codes), using the original sentence S and the semantic role set R as input to generate the target sentence S′ containing the emphatic “是” (shì).

Generation Constraints:

The emphasis function is defined as follows:

$$\text{EmphasisOp}(\Pi,\text{type} \begin{cases} \text{是} + \text{An} & \text{if type} \neq \text{ARG1} \\ \text{的是} + \text{An} & \text{if type} = \text{ARG1} \end{cases}$$

In this algorithmic formulation:

A_n represents the default syntactic position of the semantic role Π.

If the semantic role Π is a patient (ARG1), the “的是” construction is used for emphasis.

Conditional Generation Formula:

$$S' = \text{Transformer-Generate}(S, R, \text{EmphasisOp})$$

S is the original input sentence;

R is the structured representation produced by semantic role labeling (SRL) on S;

each element contains the semantic role content Π_i, the role type $type_i$ (e.g., ARG0, ARG1, ARGM-TMP), and its position span $span_i$ in the sentence;

EmphasisOp is a predefined emphasis-operation rule:

if the role type is not ARG1, the structure “是 + default position A_n” (shì + A_n) is used; if it is ARG1 (patient), the structure “的是 + A_n” (de shì + A_n) is used.

The entire generation process is implemented using a Transformer-based model (e.g., BART or T5): the model takes S and R as semantic constraints and also receives

EmphasisOp as a control condition, ultimately outputting the new emphasized sentence S′. This enables controllable natural language generation, allowing the emphasized focus and syntactic form to match the preset semantic role types precisely.

This formula is equivalent to:

$$P(S'|S, R, \text{EmphasisOp}) = \prod_{i=1}^{T} P(w_i|w_{<i}, S, R, \text{EmphasisOp})$$

3.3 Unified Algorithmic Formulation

We now formalize the entire process as an end-to-end conditional generation problem based on maximum likelihood estimation:

$$\max_{\theta} \sum_{(S,S')\in D} \log P_\theta(S'|S, R(S), \text{EmphasisOp})$$

This formula frames the generation of emphasis sentences guided by semantic roles as an end-to-end conditional maximum likelihood estimation problem:

Training set D={(S,S′)}: each sample contains an original sentence S and a manually annotated emphasized sentence S′.

Semantic role extraction R(S): a set of semantic roles automatically extracted from S by the BERT-SRL model, with each role containing content, type, and position information.

Emphasis operation rule EmphasisOp: selects the syntactic template “是 + A_n” (shì + A_n) or “的是 + A_n” (de shì + A_n) according to the role type.

Generation model Pθ: a Transformer parameterized by θ, where the conditional probability Pθ(S′|S,R(S),EmphasisOp) represents the likelihood of generating the target sentence given the original sentence, the SRL structure, and the operation rule.

Based on the above analysis, we propose a two-stage deep generative model:

R=BERT-SRL(*S*)

S′=Transformer-Generate(*S, R,* EmphasisOp)

The EmphasisOp function determines whether to insert “是” or “的是” for emphasis in the syntactic structure according to the type of semantic role in the event structure.

4. SUPERVISED LEARNING AND EXPERIMENTAL ANALYSIS

4.1 Corpus and Training Set Design

Supervised learning constructs models from training data to enable accurate predictions on unseen data (Müller & Guido 2018). This study designs a parallel corpus of 1,000 sentences focusing on the emphatic “是” in Chinese, comprising training, validation, and test sets. The corpus is drawn from modern Chinese novels, news texts, and grammatical examples to ensure diversity and representativeness.

During data preprocessing, we implement a dual-validation mechanism that combines automatic annotation with manual verification. First, existing Chinese Semantic Role Labeling (SRL) tools are used for initial automatic annotation, enabling coarse-grained processing of large-scale data. Subsequently, manual fine-grained validation and correction are performed to ensure the accuracy of semantic role boundaries and the appropriateness of type determinations.

The parallel corpus serves as annotated data for supervised learning, ensuring that the model can accurately learn the syntactic generation patterns of the emphatic “是”. The final dataset contains 1,000 sentence pairs, divided into training, validation, and test sets in a 3:1:1 ratio.

Table 1. Corpus Statistics (1,000-Sentence Version)

Set	Sentence Pairs	Avg. Length	Role Distribution
Training	600	17.9 words	Uniform
Validation	200	18.2 words	Uniform
Test	200	18.1 words	Uniform

4.2 Baseline Models and Evaluation Metrics

4.2.1 Baseline Models

This study selects four representative generation methods as baselines:

- (1) Rule-Based: Uses rule-based templates to ensure syntactic correctness.
- (2) Seq2Seq+Attention: Employs a classic encoder-decoder architecture for end-to-end transformation.
- (3) GPT-2 Fine-Tuned: Leverages the generative capacity of a pre-trained language model through transfer learning.
- (4) BART: Uses a denoising autoencoder-based sequence-to-sequence model suitable for text reconstruction tasks.

4.2.2 Automated Evaluation Metrics

A multidimensional evaluation framework combining automatic and human assessments is adopted. The automatic metrics include:

- (1) BLEU-4: Evaluates the quality of generated text from the perspective of n-gram matching.
- (2) ROUGE-L: Measures sentence-level semantic coherence based on the longest common subsequence.
- (3) Semantic Role Preservation Accuracy (SRP): Specifically evaluates the accuracy with which emphasized semantic roles are preserved.

4.3 Experimental Results and Analysis

To comprehensively evaluate the performance of the proposed T5-based two-stage generation model, we conducted a systematic comparative analysis of five models across 200 test samples. The evaluation was conducted along three dimensions: BLEU-4, ROUGE-L, and Semantic Role Preservation Accuracy (SRP), thereby demonstrating the performance of each model on automatic evaluation metrics.

Table 2. Comparison of automatic evaluation results across models (200-sentence test set)

Model	BLEU-4	ROUGE-L	SRP (%)
Rule-Based	63.8	67.1	99.3
Seq2Seq+Attn	61.2	59.3	66.7
GPT-2 FT	74.5	73.1	75.9
BART	77.8	76.6	79.2
Ours (T5-based)	81.2	80.1	88.4

As shown in “Table 2”, the proposed T5-based model achieves the best performance across all three automatic evaluation metrics. Among the baseline models, the Rule-Based method performs particularly well on SRP, reaching 99.3%; however, its lower BLEU-4 and ROUGE-L scores reveal the limitations of rule-based approaches in generating natural language. Overall, the T5-based model demonstrates strong and balanced performance across all three evaluation dimensions.

5. CONCLUSION

This study applies event structure theory to natural language processing by proposing and implementing a two-stage syntactic generation system that integrates data-driven methods with knowledge-guided modeling. Through the construction of a large-scale annotated corpus and the incorporation of semantic roles as conditional control signals, the proposed model effectively generates grammatically correct and semantically accurate Chinese sentences containing the emphatic “是”.

This research has significant implications for computational linguistics, Chinese information processing, and language pedagogy. Theoretically, it validates the applicability of event structure theory in computational modeling. Practically, it provides technical insights for developing more intelligent Chinese grammar-checking systems, writing assistance tools, and pedagogical software for second language acquisition. As deep learning techniques and linguistic theory continue to converge, syntax generation based on structural understanding will play an increasingly important role in natural language processing.

ACKNOWLEDGMENTS

Fund Project: Humanities and Social Sciences Project of the Ministry of Education (23YJC740023);

The Humanities and Social Sciences Project of Henan Province's Universities "Development and Application of International Chinese Teaching Vocabulary under the Background of Building Education Strong Country Resources" (2026-ZZJH-382).

REFERENCES

[1] Fang Mei. Syntactic Realization of Contrastive Focus in Chinese. Zhongguo Yuwen (Studies of the Chinese Language), 1995. No. 4,279-288.

[2] Jin Lixin. On the Structures of “shi...de”, “shi” and “de”. Oriental Linguistics, 2013. Vol 13. 52-65.

[3] Wang Li. Chinese Modern Grammar. Beijing: The Commercial Press. 2011.

[4] Wang Ping, Shi Feng, Xiong Jinjin and Shang Sang. The Prosodic Realization of Focus in Chinese “shi” Sentences. Applied Linguistics, 2019. No. 3.134-143.

[5] Zhang, Heyou. The Semantic Function of “Shi” in the Modal-Affirmative “Shi” Construction. Journal of Peking University(Philosophy and Social Sciences), 2007. No. 2.95-101.

[6] Zhou, Guozheng. The True Identity of “Shi”: An Interpretive Marker—A Comprehensive Interpretation of the Syntactic, Semantic, and Pragmatic Functions of “Shi”. Linguistic Research, 2008. No. 2.5-11.

[7] Andreas C. Müller, Sarah Guido, Introduction to Machine Learning with Python, O’Reilly Media, Inc. and Posts &Telecom Press. 2018.