Encoding Logical Relations of Chinese Complex Sentences within the Universal Dependencies Framework

Hongpu Zhu, Hongzhi Xu
Shanghai International Studies University
{zhuhp, hxu}@shisu.edu.cn

Abstract

Clauses in complex sentences always entail certain logical relations such as conjunctive, causative, and concessive. Such logical relations, however, are not properly represented in the universal dependencies (UD) framework, being collapsed into an adverbial clause (advcl) or clausal complement (ccomp) relation between clausal heads. This study extends the UD framework by encoding 13 logical relations. With the new framework, which is structurally identical to UD, we constructed a training corpus containing about 1,769 sentences extracted from Chinese newswire and annotated an existing Chinese corpus (GSD-simp test) in UD as a test set. We trained a BERT-based biaffine parser and fine-tuned the Qwen-3 model with the training corpus and evaluated the models on the UD test data. They are compared against four general purpose LLMs including GPT-4o, GPT-5, Claude 4 and DeepSeek V3.2. We find that the fine-tuned Qwen-3-8B model achieves a UAS/LAS of 0.840/0.757, higher than the BERT-based parser and the general purpose LLMs. The results confirm the feasibility of our framework and highlight the inherent challenges of parsing hierarchical and implicit inter-clause relations.

Keywords: Chinese complex sentences; logical relation; dependency parsing; universal dependencies

1. Introduction

Understanding the logical relations of clauses in complex sentences is a key task in natural language processing (NLP), especially for computational models of semantics and discourse (Marcu, 2000; Li et al., 2022). This task is particularly challenging for Chinese, because its complex sentences often consist of several clauses juxtaposed together without overt connectives (Huang and Shi, 2016; Liang, 2002). These clauses often form a hierarchical structure, where subsequent clauses progressively add new information by incorporating preceding clauses as embedded substructures (e.g., a causative relation nested within a concessive one). Unlike English, where logical relations are typically conveyed with explicit conjunctions such as *and, or,* and *but*, inter-clause relations in Chinese are frequently implicit and ambiguous. Implicit relations can account for up to 80% of discourse relations, as reported in a treebank (Zhou and Xue, 2012). Therefore, interpretation of such logical relations requires both analysis of the syntactic structure and understanding of semantics and contextual information, posing significant challenges to computational models.

Most prior research on inter-clause logical relations has treated it as a classification task that takes two clauses as input (Sun et al., 2019; Yang et al., 2022). This approach, however, does not generalize to more common scenarios where the hierarchical organization of clauses and inter-clause logical relations are mutually dependent. A pipeline that first recursively reconstructs the clause organization and then classifies their relations in a pairwise manner would be rendered infeasible. Moreover, this interdependency suggests that they should be addressed as a unified task where the two sub tasks can draw useful information from each other.

This study addresses this issue by proposing a unified framework of structure construction and logical relation identification. After carefully examining the linguistic nature of the two, we find they can be naturally encoded in the dependency structure of sentences under e.g. the UD framework. In particular, logical relations between clauses can be represented as finer-grained dependency types between their heads, while other dependencies remain unchanged. The proposed framework has two important advantages: 1) it is structurally compatible with the existing UD framework; 2) it allows us to simultaneously model the two sub tasks that heavily rely on each other as discussed above.

This is an example sentence with three clauses, expressing two different logical relations:

(1) 陈力仁 先生 一直
Chén-Lìrén xiānshēng yīzhí
Chen-Liren Mr. always

对 农产品 非常
duì nóngchǎnpǐn fēicháng
to agricultural.products very

感兴趣， 但 由于 当地 土地
gǎn-xìngqù, dàn yóuyú dāngdì tǔdì
feel.interest, but due.to local land

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11620–11630
11–16 May 2026.

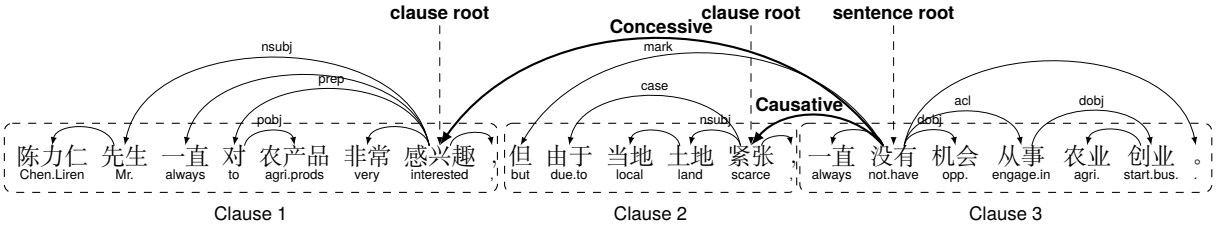

Figure 1: Full dependency graph of the example sentence under our framework. Clause roots and inter-clause relations are marked. Some dependency relations (puct, nmod) are not labeled for visual clarity.

紧张，　一直　没有　机会
jǐnzhāng, yīzhí méiyǒu jīhuì
scarce, always not.have opportunity
从事　农业　创业。
cóngshì nóngyè chuàngyè.
engage.in agriculture start.business.

"Although Mr. Chen Li-Ren has always been very interested in agricultural products, because local land is scarce, he has never had the opportunity to start a business in agriculture."

In standard UD, the word but is typically treated as a conjunction, which would make the first clause the main syntactic root. However, after inspecting the Chinese UD treebank, the equivalent connective pair " 虽然... 但是..." (*suiran...danshi*) is instead labeled as `case` and `mark`, with the second clause as the main clause. Here the first clause omits the connective " 虽然" *suiran* (although) and forms a concessive relation with the third clause. Therefore, in our framework, we treat the final clause as the main clause. The second clause provides a reason, and the first clause serves as a concessive clause modifying the main clause. Figure 1 shows the dependency graph with labeled clause relations. In bracketed form, this can be represented as:

(2) . [ROOT clause 3 [CONCESSIVE clause 1] [CAUSATIVE clause 2]]

We collected 1,769 complex sentences from Chinese news text and annotated them with the proposed framework. For benchmarking, we further annotated the GSD-simp test set (Qi, Peng and Yasuoka, Koichi, 2022) from the Chinese UD Treebank, which includes both simple and complex sentences. We established baseline performance using two models: a BERT encoder with a biaffine parser and a Large Language Model (LLM), Qwen-3. They represent two different approaches within our framework: the parser predicts token-level dependencies (like Figure 1) while LLMs output clause-level relations as a tree (as in (2), but in dependency tuples) assuming clauses are given. For comparison, we also report the performance of state-of-the-art general-purpose LLMs such as GPT-5.

The remainder of this paper is organized as follows. Section 2 reviews related work on Chinese complex sentences and similar frameworks such as discourse parsing. Section 3 introduces our annotation framework and dataset. Section 4 describes the baseline models, and section 5 presents experimental results followed by an error analysis. Finally, we conclude with a summary of contributions and limitations.

2. Related Work

The study of Chinese complex sentences has long been a central topic in Chinese linguistics. Despite rich theoretical research on their structure, computational models of inter-clause logical relations have remained relatively simple. Previous studies have primarily approached the problem as a classification task, often focusing on sentences containing only two clauses. For instance, Sun et al. (2019), Yang et al. (2022), Yang et al. (2017), and Tian et al. (2019) employed neural architectures such as CNNs, GRUs, and Transformers to classify clause pairs. Some of these studies drew data from the Chinese Complex Sentence (CCS) corpus, which is not publicly available, and adopted a limited tripartite system of coordination, causal, and transitional relations. Previous research has largely relied on punctuation for clause segmentation. For example, Xue and Yang (2011) used a maximum-entropy classifier to predict whether commas signal clause boundaries.

Beyond sentence-level, substantial efforts have been devoted to the construction of Chinese discourse treebanks. They aim to capture discourse relations between larger textual spans, e.g. sentences, paragraphs, rather than inter-clause relations within a single sentence. Two major theoretical frameworks are Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the Penn Discourse Treebank (PDTB) framework (Robaldo, 2008; Prasad et al., 2017).

RST models discourse as a hierarchical tree composed of elementary discourse units (EDUs).

EDUs are defined operationally and do not necessarily correspond to clauses. The flexibility of EDU segmentation enables RST to model discourse coherence beyond sentence boundaries. For instance, Jiang et al. (2018) constructed a macro-level treebank covering paragraph-level relations. Cao et al. (2017) discussed multiple levels of EDU segmentation, from clauses, sentences, to paragraphs. Peng et al. (2022) introduced discontinuous EDUs to capture embedded structures such as relative clauses. Although similar, the RST framework cannot be easily applied to clauses within complex sentences. Some types in its taxonomy go beyond the logical relations addressed in this study, such as elaboration, summary, and background.

In contrast, PDTB relies on explicit connectives to define discourse relations. The connective serves as a predicate and connects exactly two arguments. When an explicit connective is absent, annotators need to insert an appropriate connective. If this is not possible, annotators would specify whether the relation is realized through a lexicalized content word (AltLex), describing the same entity (EntRel), or absent altogether (NoRel).

As noted by Zhou and Xue (2012), this framework is not fully suitable for Chinese, where implicit relations can account for up to 80% of all discourse relations. Under its constraints, some relation types cannot be annotated despite being clearly definable. The CUHK Discourse Treebank (Zhou et al., 2014) and the Chinese Discourse Treebank (Zhou and Xue, 2015) attempted to adapt PDTB conventions by using punctuation as the connective. Moreover, PDTB does not organize multiple logical relations in one sentence within a hierarchical tree structure. Although PDTB provides a fine-grained taxonomy of relation senses, its structural design is insufficient for modeling Chinese complex sentences.

Another relevant task is the Hierarchical Clause Annotation (HCA) framework for English proposed by Fan et al. (2023). HCA segments complex sentences into clauses and annotates their interrelations in a tree structure, using both constituent and dependency representations during annotation. However, it was designed primarily for Abstract Meaning Representation (AMR) parsing and focuses on syntactic or semantic relations such as predicative, appositive, or relative clauses rather than logical ones. Also, it requires a more complicated top-down parsing process that differs from standard approaches of dependency parsers.

Beyond these, Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2008) provides a prominent dynamic semantic theory of discourse interpretation, using rhetorical relations to model the semantics/pragmatics interface. SDRT relates semantically underspecified forms generated by a grammar to their pragmatically preferred interpretations using a glue logic based on commonsense reasoning. Despite its theoretical robustness and explanatory power, SDRT relies heavily on logical forms of clauses, which cannot be easily parsed and represented.

In summary, previous research on Chinese complex sentences and discourse annotation provides valuable reference for the current study. However, the discourse framework cannot be freely extended to model logical relations between inner-sentence clauses. The framework proposed in this paper complements the discourse-level resources by targeting logical relations between clauses within Chinese complex sentences and can be integrated together to model the overall logical structure across different spans of text.

3. Annotation framework

3.1. Encoding Clausal Relations in Dependency Structures

Clauses in a Chinese complex sentence inherently entail hierarchical logical relations between them. These relations can be naturally represented with a dependency tree structure, where each node can either be a simple clause, or a subtree consisting of embedded complex clauses. The tree structure is fully determined by semantics of individual clauses and the logical relations between them. This hierarchical property aligns well with dependency grammar, since the logical relations between clauses can be encoded through the dependency relations connecting their predicate heads. It is thus straightforward to extend the existing dependency framework, e.g. the UD framework by defining finer-grained dependency types between heads of clauses.

In the current UD framework, dependency relations between clause predicates are typically labeled with four categories: `advcl`, `ccomp`, `xcomp`, and `parataxis`. The relation `advcl` (adverbial clause modifier) links a subordinate clause functioning as an adverbial modifier of the main clause. The majority of clause predicates in our corpus fall under this relation, including those of causal, conditional, concessive, and purposive clauses. The relations `ccomp` (clausal complement) and `xcomp` (open clausal complement) also appear frequently, because many clauses appear as complements of prepositional or verbal connectives such as 由于 (“due to”) and 为了 (“in order to”). The fourth relation, parataxis, covers juxtaposed or coordinated clauses of roughly equal status, often without explicit connectives.

Although these four types adequately capture

Major Category	Subtype	Definition	Example Sentence
Conjunctive (Both statements share the same truth value.)	Equivalent	Clauses of equal status, expressing parallel information. The order can often be interchanged.	他喜欢看书，也喜欢旅游。(He likes reading, and also likes traveling.)
	Progressive	The second clause adds stronger or further meaning.	他很努力，甚至每天学习到午夜。(He is hard-working, even studies until midnight.)
	Temporal	Clauses express sequential events.	他吃了饭，然后去上班。(He ate, then went to work.)
	Comparative	Clauses indicate comparison.	他跑得越快，我越开心。(The faster he runs, the happier I am.)
Disjunctive (Only one statement is true)	Equivalent	Clauses present symmetrical alternatives.	你可以去北京，或者去上海。(You can go to Beijing, or go to Shanghai.)
	Non-equivalent	One alternative is preferred over the other.	与其成为无名小卒，不如努力扬名立万。(Rather than live as Nobody, work hard and go down in times in glory.)
Causative	–	One clause provides the cause of another.	因为下雨，比赛取消了。(Because it rained, the game was cancelled.)
Conditional	Sufficient	The condition guarantees the outcome.	如果下雨，就取消比赛。(If it rains, the game will be cancelled.)
	Necessary	The condition must hold for the outcome.	只有努力，才会成功。(Only if you work hard, you will succeed.)
	Hypothetical	Expresses hypothetical or imagined condition.	假如我是你，我会去。(If I were you, I would go.)
	Exhaustive	The main clause is true regardless of the condition.	无论谁来，我都欢迎。(No matter who comes, I will welcome them.)
Concessive	–	A clause is true despite another clause.	虽然下雨，大家仍然出门。(Although it rained, everyone still went out.)
Purposive	–	One clause expresses the purpose of another.	他努力学习，为了通过考试。(He studies hard in order to pass the exam.)
ROOT	–	Root of a hierarchical clause tree.	–

Table 1: Relation taxonomy in our annotation schema. Each relation is illustrated with a definition and a simple example sentence.

syntactic configurations of clausal dependency, this mapping between syntactic dependency and logical relation is not one-to-one. Clauses with the same inter-clause logical relations can appear under different dependency labels. Our framework therefore provides a finer-grained relation labels between clause predicates according to their logical functions, such as causative, concessive, or conditional, thereby naturally capturing the logical relations along with the hierarchy structure within existing UD framework.

In UD, the dependency arcs typically span from a modifying clause to the modified clause. However, the UD standard does not provide any specific rules for determining the attachment direction between clauses with different logical relations (Leung et al., 2016; Poiret et al., 2023), and inconsistencies can be observed in the GSD-simp treebank. We adopt a uniform convention that the clause carrying the most semantically prominent information, typically the final clause, is assigned as the parent. When two clauses have equal semantic and syntactic status, the first clause is assigned as the parent.

Figure 2 further illustrates the hierarchical relations. Such a hierarchical clause tree can be easily constructed with a traversal of arcs between clause predicates. These two tree structures (a) and (b) have different interpretations. In both of them, the first three clauses together form a progressive relation with the final clause. In (a), since clause 1 attaches to clause 3, it has the interpretation of providing a reason for the progressive sequence of clause 2 and 3. However, in (b), clause 1 attaches to clause 2, and provides a reason for only clause 2. Clause 1 and 2 together form a progressive relation with clause 3.

3.2. Taxonomy of Logical Relations

For annotating inter-clause logical relations, we follow the taxonomy defined in *A Reference Grammar of Chinese* Huang and Shi (2016). The framework distinguishes six major categories and thirteen subcategories of logical relations, as sum-

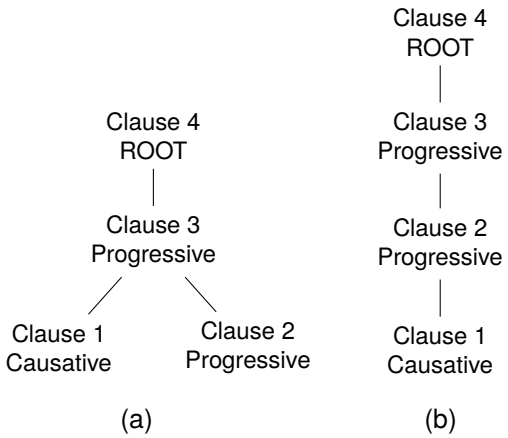

Figure 2: Two alternative hierarchical structures of logical relations: (a) Clause 1 attaches under Clause 3; (b) Clause 1 attaches under Clause 2.

marized in Table 1. The six major categories include conjunctive, disjunctive, causative, conditional, concessive, and purposive relations. In both conjunctive and disjunctive relations, two clauses A and B are syntactically equal and relatively independent. However, conjunction requires that both statements be true or false at the same time, while disjunction does not. In particular, for concessive relations, the truth of the adjunct clause seems to indicate the main clause will be false, but it actually does not detract from its truth Huang and Shi (2016). Therefore, the typical *suiran... danshi* ("although... but...") structures are categorized as concessive instead of conjunctive. In addition, the label ROOT is used for the root clause of the sentence, typically the clause that contains the root token of the whole sentence.

3.3. Descriptive statistics

The annotated training corpus comprises 1,769 manually annotated Chinese complex sentences. These sentences are drawn from the Chinese Gigaword corpus (Parker, Robert et al., 2011) and the GSD-simp training set. Each sentence is annotated with (a) clause boundaries, (b) the root token of each clause, and (c) logical relation labels between clauses. The resulting dataset contains approximately 6,500 clauses, with an average of 3.68 clauses per sentence. For benchmarking, the GSD-simp test set is annotated using the same scheme and includes 500 simple and complex sentences.

Table 2 summarizes the distribution of logical relation types, and Table 3 shows the five most frequent dependency relations among clause root tokens (excluding the ROOT labels). These distributions demonstrate that our annotation essentially provides a finer-grained classification of clause-level dependencies, such as `advcl` and `ccomp`.

Major Category	Subtype	Train	Test
Conjunctive	Equivalent	290	58
Conjunctive	Progressive	1,275	197
Conjunctive	Comparative	77	2
Conjunctive	Temporal	206	57
Disjunctive	Non-equivalent	124	2
Disjunctive	Equivalent	64	1
Conditional	Sufficient	158	4
Conditional	Necessary	177	1
Conditional	Hypothetical	183	4
Conditional	Exhaustive	112	1
Concessive	Concessive	272	31
Purposive	Purposive	178	14
Causative	Causative	592	89
ROOT	–	1,769	500

Table 2: Distribution of inter-clause relations in the annotated corpus.

Relation Type	Count	Percentage
advcl	1,280	34.5%
ccomp	709	19.1%
parataxis	617	16.6%
xcomp	602	16.2%
csubj	72	1.9%

Table 3: Distribution of major dependency relations of clause root tokens in the training dataset.

3.4. Annotation Procedure and Quality control

The primary annotation was performed by one native Chinese speaker with knowledge of Chinese linguistics. Detailed documentation of the annotation schema was discussed and agreed upon prior to the task. A simple web-based interface was developed for the process. It displayed the whole sentence, as well as dependency parsing and constituent parsing results as reference. The annotator then adjusted the dependency arcs between clause root tokens and corrected them if needed, according to their understanding of the sentence.

A subset of 200 sentences was independently annotated by two annotators, and inter-annotator agreement was measured using accuracy and Cohen's κ, as shown in Table 4. Given the high agreement, the remaining sentences were annotated and subsequently reviewed again by one annotator.

Item	Cohen's κ	Acc,
Clause head	0.965	0.975
Inter-clause relation	0.941	0.953

Table 4: Inter-rater agreement for clause head and inter-clause relation.

4. Experiments

For the task, we trained two baseline models: a BERT-based biaffine parser and a LLM with Low-rank Adaptation (LoRA) fine-tuning, representing two distinct approaches: token-level dependency parsing and clause-level structure generation.

The first biaffine dependency parser is built with a BERT encoder (Dozat and Manning, 2017) following the UDify architecture (Kondratyuk and Straka, 2019). It is implemented using HanLP (He and Choi, 2021). The encoder is initialized with `bert-base-chinese` and then trained on the GSD training set of the Chinese UD Treebank as a standard dependency parser. The model achieves state-of-the-art performance on the test set (UAS: 0.876, LAS: 0.845). The parser is then further fine-tuned on our training set to learn to parse inter-clause arc labels representing logical relations between clauses.

The second system uses Qwen-3 (Yang et al., 2025), a decoder-only language model. In this setup, the task is reformulated as text generation, and the model predicts a dependency tree for the clauses directly. The clause tree is programmatically constructed from the annotated token-level dependency corpus. The full annotation guidelines and relation schema, along with examples are provided in the system prompt, and each input sentence is presented as the user prompt. The model is trained to generate an array of dependency tuples in which each element specifies the head index and logical relation label of a clause. Four Qwen-3 variants (0.6B, 1.7B, 4B, and 8B parameters) are fine-tuned with LoRA for three epochs under identical hyperparameter settings. These models are also trained on our 1,769-sentence training set. For evaluation, they were tested on the annotated 500-sentence GSD-simp test set, where gold clause segmentation was provided to isolate the models' performance on logical relation prediction.

Several state-of-the-art LLMs are also evaluated for their few-shots performance on the task. They include Claude 4 Sonnet, GPT-4o, GPT-5, and DeepSeek V3.2. The testing procedure, including prompts and generation format, is the same as the Qwen-3 model.

Performance is assessed using standard dependency metrics, unlabeled attachment score (UAS) and labeled attachment score (LAS). We also report accuracy and macro F1 score for logical relation classification. For the BERT model, performance is measured on the clause root tokens, and for LLM systems, measured at the clause level.

5. Results and discussion

5.1. Model performance

Table 5 summarizes the overall performance of all models on the test set. The BERT-based parser has an overall UAS and LAS of 0.833 and 0.785 on the test set. On clause root tokens, this drops to 0.693 and 0.583. Among all systems, the LoRA-tuned Qwen-3 8B model achieves the highest scores across all metrics, (UAS = 0.840, LAS = 0.757, relation accuracy = 0.790, macro F1 = 0.546), outperforming both the BERT-based parser and other Qwen variants. When relations are grouped by major categories, accuracy and macro F1 rise to 0.828, and 0.656 respectively. Detailed classification performance of each relation type is reported in Table 6. Other Qwen models also achieve competitive results, and the BERT-based biaffine parser shows lower UAS and LAS than all Qwen models. Its overall performance is comparable to general-purpose LLMs such as GPT-4o. All general-purpose LLMs perform substantially below the fine-tuned Qwen models, indicating that decoder-only architectures without fine-tuning are less effective for tasks requiring explicit structural prediction such as ours. GPT-5 yields the strongest results among these LLMs, especially in terms of UAS.

Model	UAS	LAS	Acc.	f1-m
bert-base-ch	0.693	0.583	0.634	0.497
Qwen 3-0.6 B	0.793	0.695	0.728	0.415
Qwen 3-1.7 B	0.817	0.736	0.766	0.431
Qwen 3-4 B	0.820	0.746	0.775	**0.576**
Qwen 3-8 B	**0.840**	**0.757**	**0.790**	0.546
GPT-4o-latest	0.632	0.532	0.589	0.386
GPT-5	0.784	0.618	0.657	0.339
Claude 4 Sonnet	0.695	0.621	0.645	0.430
Deepseek V3.2	0.692	0.563	0.608	0.411

Table 5: Overall model performance on the test set. For the BERT model, performance is measured on the clause root tokens, and for LLM systems, measured at the clause level.

Figure 3 illustrates how parsing performance (UAS) varies with sentence complexity, measured by the number of clauses per sentence. All models perform well on simple sentences containing one or two clauses, but accuracy declines rapidly as the number of clauses increases. The BERT parser shows the steepest degradation, dropping to 0.35 UAS on sentences with five or more clauses. In contrast, the LoRA-tuned Qwen models maintain greater robustness. Qwen-4B achieves 0.81 UAS on three-clause sentences and remains above 0.6 for more complex inputs. Larger Qwen variants (4B–8B) consistently outperform smaller ones, indicating that greater model

Category	P	R	F1
Conjunctive-Equivalent	0.517	0.588	0.551
Conjunctive-Progressive	0.731	0.670	0.699
Conjunctive-Comparative	1.000	0.333	0.500
Conjunctive-Temporal	0.807	0.708	0.754
Disjunctive-Non-equivalent	0.500	0.500	0.500
Disjunctive-Equivalent	0.000	0.000	0.000
Conditional-Sufficient	0.500	0.500	0.500
Conditional-Necessary	0.000	0.000	0.000
Conditional-Hypothetical	0.500	0.400	0.444
Conditional-Exhaustive	1.000	1.000	1.000
Concessive-Concessive	0.645	0.769	0.702
Purposive-Purposive	0.429	0.546	0.480
Causative-Causative	0.596	0.639	0.616
ROOT	0.902	0.912	0.907
Accuracy		0.790	
Macro avg	0.580	0.540	0.547
Weighted avg	0.796	0.790	0.792

Table 6: Precision, recall, and F1-scores for each clause relation category on the test set of Qwen3-8B.

capacity improves generalization to longer and nested structures. Overall, the results demonstrate that inter-clause dependency prediction becomes increasingly challenging as clause count rises. This seems to only affect clause-level parsing, as performance of BERT parser over all tokens only exhibits modest decline.

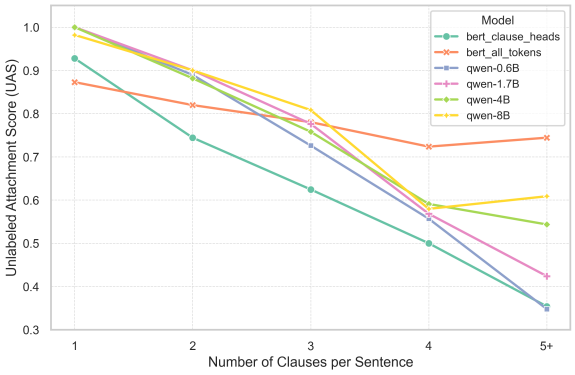

Figure 3: Clause-level parsing accuracy by number of clauses per sentence.

5.2. Error analysis

To better understand model behavior, we conducted a clause-level error analysis on the BERT-based biaffine parser (Table 7). We distinguish three types of errors: (1) incorrect identification of a clause root token, (2) incorrect attachment to a parent clause, and (3) incorrect labeling of inter-clause relations. Errors are counted at the clause level. Among 1,077 test clauses, 70.8% have the correct parent clause and 70.3% have the correct relation label, when attachment and labeling are evaluated by traversing the predicted dependency tree with depth-first search, allowing for misidentification of clause root tokens. Only 55.1% of clauses are correctly predicted across all three dimensions, consistent with the token-level LAS reported in Table 5. We also observe 66 cases in which the parser incorrectly assigns a logical relation label to a non-root token.

Condition	Count	Percentage (%)
Correct root + parent + label	593	55.1
Parent + label correct	644	59.8
Root + label correct	694	64.4
Root + parent correct	694	64.4
Root correct only	961	89.2
Parent correct only	762	70.8
Label correct only	757	70.3
Total clauses	**1,077**	

Table 7: Clause-level error analysis of root, parent, and relation label prediction. Percentages are relative to the total number of test clauses (1,077).

In comparison with the identification of logical relations between clauses, we investigated the performance of the original UD parser (without fine-tuning for logical relations) on identifying dependencies between roots of clauses. The performance is substantially lower (UAS 0.734, LAS 0.695) than the parser's overall performance (UAS 0.876, LAS 0.845), indicating that dependency relations between clauses are inherently more difficult to capture. This also shows that our fine-tuned parser has a very promising performance on identifying logical relations between clauses (UAS 0.840, LAS 0.757).

Because the UD test set contains an imbalanced distribution of relation types and many implicit or semantically ambiguous relations, we selected an additional subset comprising ten sentences with only two clauses for each logical relation, all featuring explicit and unambiguous connectives. This is used exclusively to assess relation classification accuracy and to evaluate the impact of marked versus unmarked relations. The models perform substantially better on this marked set compared to the general test set, confirming that the absence of overt connectives (implicit relations) is a primary source of parsing errors.

The resulting confusion matrix (Figure 4) indicates that causative, concessive, and conjunctive–comparative relations are recognized with high accuracy. Conditional subtypes are also reasonably well distinguished, though moderate cross-prediction occurs among them, and many hypothetical conditions are classified as sufficient conditions. In contrast, conjunctive–equivalent and disjunctive–equivalent relations are rarely identified correctly, with many instances misclassified

as progressive clauses or not identified at all. The model also shows a systematic bias toward the progressive category, likely reflecting its higher frequency in the training data. Certain types, such as Conjunctive-Comparative, which have rather obvious connectives that are often not omitted, showed better performance.

The poor performance on disjunctive-equivalent relations may seem surprising given the existence of explicit disjunctive markers like *huozhe* ("or"). However, unlike English where "or" reliably separates clauses, Chinese disjunctive markers are rarely used for connecting full clauses in the dataset. When Chinese complex sentences do express clausal disjunction, they often rely on parallel syntactic structures without overt connectives. Furthermore, structurally disjunctive clauses might seem identical to a conjunctive one, and differentiating the two requires the model to evaluate the truth value of two statements, whether they are mutually exclusive or they are both true. This structural ambiguity, combined with a severe class imbalance (accounting for only 64 instances in the training set), causes the model to struggle.

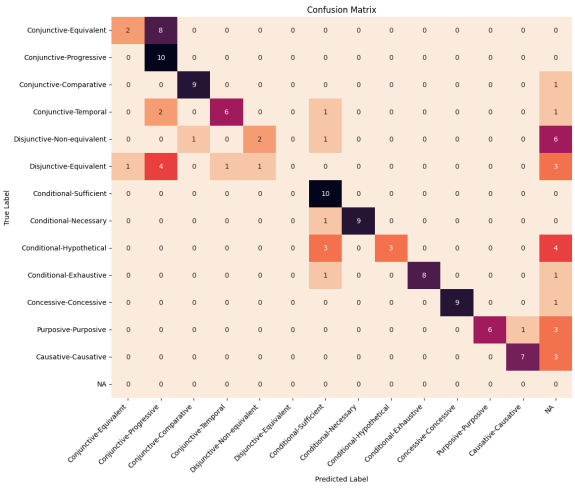

Figure 4: Confusion matrix on the extra diagnostic examples.

5.3. Discussion

The study is related to but different from discourse parsing. While discourse parsing typically emphasizes logical relations across broader textual spans, this study is specifically focused on dependency-based logical relations between clauses within individual complex sentences. They complement each other and can be combined to represent the whole logical structure of text. The baseline models we trained are comparable to those reported in previous studies on Chinese discourse parsing. For instance, Cheng and Li (2019) developed a dependency-based system with transfer learning for RST-style discourse parsing and reported UAS/LAS scores of 0.64 and 0.38, while Peng et al. (2022) achieved F1-scores of approximately 0.5–0.7 on the GCDT treebank using several BERT-based parsers. With the proposed framework, it is shown that the LoRA-fine-tuned Qwen-3 8B model achieved the best overall performance (UAS 0.840, LAS 0.757), outperforming both the BERT-based parser and general-purpose LLMs such as GPT-4o. This suggests that our annotation framework provides a valid and practical foundation for modeling clause-level logical relations within sentences.

The BERT-based parser, though performs well on standard UD parsing, shows a significant decline in performance when predicting inter-clause relations. Further investigation of parser behavior reveals the task's inherent difficulty. UAS and LAS scores on clause-root tokens remain well below average even in standard UD parsing (0.734/0.695 compared with 0.876/0.845 overall). This reflects the challenges brought by longer dependency distances and a mismatch between syntactic dependency and discourse relations. The fact that the accuracy of all models systematically declines as clause count increases further supports the view that clause-level logical parsing introduces additional structural complexity beyond standard syntactic parsing or classification.

Furthermore, the error analysis reveals frequent confusion among categories that require deeper semantic understanding than surface features can provide. For instance, the model often confuses different types of conditional relations and struggles to distinguish disjunctive from conjunctive relations. This demonstrates that the task requires a higher level of competence than standard syntactic parsing. Correctly distinguishing these relations often requires reasoning capabilities, such as causality inference, condition assessment and evaluation of propositions.

The imbalanced training set and the prevalence of implicit relations in Chinese have an adverse effect on model performance. Moreover, the current corpus lacks genre diversity, as most sentences are drawn from newswire and formal texts. Expanding the dataset to include a broader range of registers, such as academic writing or fictional writing can enhance model generalization. Building larger and more varied annotated corpora thus represents a crucial next step toward developing models capable of deeper and more comprehensive language understanding.

6. Conclusion

This paper introduced a dependency-based framework for parsing clause-level logical relations, aimed at addressing the specific challenges of Chinese complex sentences. Building on the UD framework, our approach represents inter-clause relations as labeled dependency arcs between the predicate heads of clauses and organizes multiple logical relations within a sentence into a hierarchical tree structure. We annotated both a training and a testing corpus with detailed guidelines and achieved high inter-annotator reliability.

Baseline results using a BERT-based biaffine parser and LoRA-tuned Qwen models support the validity and feasibility of the framework, but current architectures still face notable challenges with implicit or nested relations. Model performance decreases as sentence complexity increases, reflecting the difficulty of modeling long distance discourse relations within a sentence. Lower performance of general-purpose LLMs in few-shots testing further indicates that both strong language understanding and task-specific adaptation are necessary for reliable predictions of hierarchical inter-clause relations.

Overall, the framework provides an intermediate representation between syntactic parsing and discourse analysis, bridging the gap between token-level dependencies and broader discourse structures such as those in RST and PDTB. Future work will extend the corpus to more diverse text types and explore models that integrate syntactic, semantic, and discourse features for more consistent and interpretable clause-level relation parsing.

7. Ethical Considerations

All data used in this study were drawn from publicly available sources and datasets. No personal or sensitive information is included. Annotators participated voluntarily. The work complies with general research ethics and data-use guidelines.

8. Limitations

This study has several limitations. The corpus primarily consists of newswire and formal written texts, which may not capture the full diversity of clause-linking patterns found in conversational or literary Chinese. Clause segmentation occasionally involves ambiguous cases, such as embedded clauses. Implicit inter-clause relations are not discussed, and the most obvious relation type is selected. The corpus size is modest, and some relation types are underrepresented, which may affect model performance. The evaluated models were trained and tested on texts from similar domains, and their generalization to other genres or discourse styles remains to be examined.

9. Bibliographical References

Shuyuan Cao, Nianwen Xue, Iria da Cunha, Mikel Iruskieta, and Chuan Wang. 2017. Discourse Segmentation for Building a RST Chinese Treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 73–81, Santiago de Compostela, Spain. Association for Computational Linguistics.

Yi Cheng and Sujian Li. 2019. Zero-shot chinese discourse dependency parsing via cross-lingual mapping. *arXiv preprint arXiv:1911.12014*.

Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. ArXiv:1611.01734 [cs].

Yunlong Fan, Bin Li, Yikemaiti Sataer, Miao Gao, Chuanqi Shi, Siyi Cao, and Zhiqiang Gao. 2023. Hierarchical Clause Annotation: Building a Clause-Level Corpus for Semantic Parsing with Complex Sentences. *Applied Sciences*, 13(16):9412. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.

Han He and Jinho D. Choi. 2021. The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders. ArXiv:2109.06939 [cs].

Chu-Ren Huang and Dingxu Shi. 2016. *A reference grammar of Chinese*. Cambridge University Press.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. Mcdtb: a macro-level chinese discourse treebank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504.

Dan Kondratyuk and Milan Straka. 2019. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. ArXiv:1904.02099 [cs].

Alex Lascarides and Nicholas Asher. 2008. Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure. In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning*, pages 87–124. Springer Netherlands, Dordrecht.

Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing Universal Dependencies for Mandarin Chinese. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 20–29, Osaka, Japan. The COLING 2016 Organizing Committee.

Bin Li, Miao Gao, Yunlong Fan, Yikemaiti Sataer, Zhiqiang Gao, and Yaocheng Gui. 2022. DynGL-SDP: Dynamic graph learning for semantic dependency parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3994–4004.

Yunhua Liang. 2002. An analysis of the compact complex structure in modern chinese. 19(2):99–106.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3).

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing. ArXiv:2210.10449 [cs].

Rafaël Poiret, Tak-Sum Wong, John Lee, Kim Gerdes, and Herman Leung. 2023. Universal Dependencies for Mandarin Chinese. *Language Resources and Evaluation*, 57(2):673–710.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The Penn Discourse Treebank: An Annotated Corpus of Discourse Relations. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1197–1217. Springer Netherlands, Dordrecht.

Livio Robaldo. 2008. The penn discourse treebank 2.0. In *Lrec*.

Kaili Sun, Yuan Li, Dunhua Deng, and Yang Li. 2019. Multi-channel CNN based inner-attention for compound sentence relation classification. *IEEE Access*, 7:141801–141809. Publisher: IEEE.

Wenhong Tian, Yinquan Gao, Houwen Huang, Zaiwan Li, and Zhaoyang Zhang. 2019. Implicit Discourse Relation Analysis Based on Multi-task Bi-LSTM. *Journal of Chinese Information Processing*, 33(5):47–53.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 631–635, Portland, Oregon, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. ArXiv:2505.09388 [cs].

Jincai Yang, Zhongzhong Chen, Xianjun Shen, and Jinzhu Hu. 2017. Automatic recognition of relation category of non-saturated compound sentences with two clauses. *Application Research of Computers/Jisuanji Yingyong Yanjiu*, 34(10).

Jingcai Yang, Yuxin Cao, Hu Quan, and Cai Xu-Xun. 2022. Automatic Recognition of Chinese Compound Sentence Relation Based on BERT-FHAN Model and Sentence Features. *Computer Systems & Applications*, 31(9):233–240.

Lanjun Zhou, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2014. The CUHK Discourse TreeBank for Chinese: Annotating Explicit Discourse Connectives for the Chinese TreeBank. In *LREC*, pages 942–949.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–77.

Yuping Zhou and Nianwen Xue. 2015. The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

10. Language Resource References

Parker, Robert, Graff, David, Chen, Ke, Kong, Junbo, and Maeda, Kazuaki. 2011. Chinese Gigaword Fifth Edition.

Qi, Peng and Yasuoka, Koichi. 2022. *UD Chinese GSDSimp*. PID https://github.com/UniversalDependencies.

A. Data availability

The annotated dataset used for the baseline models in this study are publicly available in our GitHub repository: `https://github.com/Zhurp2020/Chinese_UD_logic`. These sentences are all compiled from the GSD-simp training/test set and the Chinese Gigaword corpus.