Open Access Library Journal
2026, Volume 13, e15785
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Corpus Construction and NER Model Comparison for Formation Name Recognition in Completion Reports

Yingru Cai

School of Foreign Languages, Xi'an Petroleum University, Xi'an, China
Email: iamyrcai@163.com

How to cite this paper: Cai, Y.R. (2026) Corpus Construction and NER Model Comparison for Formation Name Recognition in Completion Reports. *Open Access Library Journal*, **13**: e15785.
https://doi.org/10.4236/oalib.1115785

Received: July 20, 2026
Accepted: August 11, 2026
Published: August 14, 2026

Abstract

Completion reports are central to the oil and gas exploration and development process, in which formation names serve as fundamental entities for geological modeling and reserve calculation. However, general-purpose Chinese named entity recognition (NER) tools do not include formation names as an entity type, and the petroleum domain lacks Chinese corpus annotation at the sequence labeling level. Based on the World Oil Outlook 2026 (WOO 2026) published by OPEC, this study employs AntConc frequency analysis and KH Coder co-occurrence network analysis to design data-driven annotation rules, and completes independent annotation and adjudication review through Doccano. The results show that the fully-featured conditional random field (CRF) model combined with rule-based post-processing achieves an F1 score significantly superior to the rule-based baseline. This provides foundational resources and methodological references for Chinese information extraction in the petroleum domain.

Subject Areas

Linguistics

Keywords

Completion Reports, Oil and Gas, Corpus Annotation, Conditional Random Fields

1. Introduction

According to the World Oil Outlook 2026 (WOO 2026) published by OPEC, oil remains the largest single source of energy [1]. In sustained high-intensity exploration and development activities, completion reports serve as the core technical

documents recording drilling geological information, carrying key geological data such as stratigraphic division, lithological description, and well logging interpretation. Among them, formation names are fundamental entities linking geological modeling, reserve estimation, and inter-well correlation; their accurate recognition and structured extraction are of significant importance for digital oilfield construction and geological knowledge graph development. However, the automatic recognition of formation names faces multiple challenges. First, the entity type system of general-purpose NER tools is primarily oriented toward general categories such as person names, place names, and organization names, and does not include formation names as a domain-specific entity type. Second, formation names exhibit the phenomenon of concurrent use of three types of variants—full names, abbreviations, and codes—in text, which increases the difficulty of recognition. Furthermore, NER corpus annotation in the Chinese petroleum domain remains insufficient, which is unfavorable for the training and evaluation of supervised learning models.

Based on the above, this paper primarily investigates the following questions: 1) How to construct the annotated corpus for formation names in completion reports; 2) How to design data-driven annotation guidelines based on frequency analysis and co-occurrence network analysis, directly mapping linguistic feature discoveries to annotation rules; 3) To verify the effectiveness of the constructed corpus through systematic comparison between a domain CRF model and rule-based methods, along with feature ablation experiments.

2. Development and Research Status of NER Technology

2.1. NER Development Background

Named entity recognition (NER) technology has undergone an evolutionary process from statistical models to deep learning. Conditional random fields (CRF), as a traditional sequence labeling method, model the dependencies of label sequences through global normalization and can achieve good recognition performance when feature engineering is sufficient [2]. The subsequently proposed BiLSTM-CRF model utilizes bidirectional long short-term memory networks to automatically extract contextual features, reducing the reliance on manual feature selection [3]. In recent years, NER performance has been further enhanced through large-scale corpus pre-training to obtain rich linguistic knowledge representations [4]. These all belong to the sequence labeling paradigm, with CRF as the methodological starting point of this paradigm, holding a foundational position in the development of NER technology.

2.2. NER Research Status

Currently, BioBERT in the biomedical domain and LegalNER in the legal domain have achieved remarkable success by constructing domain-specific annotated corpora to drive specialized models [5]. In the petroleum domain, a limited number of information extraction studies targeting geological texts have been published

in English literature [6], primarily focusing on pattern matching for entities such as well names and depths [7]. Natural language processing research on petroleum texts has mainly focused on terminology form or basic construction. By constructing an annotated corpus for formation names in completion reports, this paper provides additional practical references for NER sequence labeling in the petroleum geology domain and simultaneously improves the research efficiency for such texts.

3. Corpus Construction and Annotation

3.1. Corpus Source and Linguistic Feature Analysis

This study selected major chapters of WOO 2026 and desensitized completion reports as the original corpus, with a total scale of approximately 120,000 characters. To systematically reveal the distribution patterns and semantic associations of formation names in text, AntConc and KH Coder were employed for frequency analysis and co-occurrence network analysis, respectively. The AntConc frequency analysis results showed that formation names frequently appear near verbs such as "drilled into", "encountered", and "drilled to", forming a significant verb-noun collocation pattern. The KH Coder co-occurrence network analysis further revealed close semantic associations between formation names and terms such as "thickness", "depth", "lithology", and "age". In the co-occurrence network, formation name nodes and the aforementioned term nodes form high-weight edge connections, constituting a semantic cluster centered on formation names with geological attribute descriptions on the periphery. This structural feature has direct guiding significance for boundary rule design.

Specifically, the corpus comprises 42 desensitized completion reports (3560 sentences) and 8 major chapters from WOO 2026 (676 sentences), totaling 50 documents and 4236 sentences. WOO chapters were included as supplementary training material because the report contains standardized petroleum terminology, geological formation descriptions, and stratigraphic references across multiple basins and regions, providing a formal technical register comparable to completion reports. Additionally, the WOO text introduces diverse linguistic contexts for formation name variants—including full names, abbreviations, and codes—that may not be fully represented in the available completion reports alone, thereby enriching the lexical and contextual coverage of the training data.

3.2. Annotation Scheme

Based on the linguistic feature analysis results, this study developed an entity annotation scheme for formation names. Nominal expressions in the report text referring to specific lithostratigraphic units are covered, encompassing four hierarchical levels: Formation, Member, Submember, and Bed. The annotation adopts the BIO tagging scheme, with the entity type labeled as "FORM" (Formation). At the morphological level, formation names follow the formation rule of "[place name/rock name] + [Formation/Member/Submember/Bed] + [sequence num-

ber]", and exhibit three types of variants: full names, abbreviations, and codes; in terms of boundary rules, complete hierarchical names are labeled as single entities; in terms of disambiguation, distinguishing between geological ages and formation names is the primary challenge. When expressions such as "Paleogene" are collocated with verbs such as "drilled into" and "encountered", they are classified as formation name entities; when collocated with verbs or nouns such as "deposition" and "depositional period", they are classified as geological ages and are not annotated.

To illustrate the BIO annotation scheme for each variant type, the following examples demonstrate the intended entity boundaries: 1) Full name: *e.g.*, "Guantao Formation", tagged as B-FORM (Guantao) I-FORM (Formation), where the entire multi-token expression is labeled as a single entity; 2) Abbreviation: *e.g.*, "Sha-3 Member", tagged as B-FORM (Sha) I-FORM (-) I-FORM (3) I-FORM (Member), where the hierarchical suffix and sequence number are included within the entity boundary; 3) Code: *e.g.*, "Ng", tagged as B-FORM (Ng), where the compact alphanumeric string is labeled as a single-token entity.

3.3. Annotation Process and Statistics

The annotation work was conducted on the Doccano platform, with independent annotation performed by master's students with backgrounds in geology and linguistics, and quality control overseen by supervisors from the same disciplines. The annotation process comprised four stages: pilot annotation, guideline revision, formal annotation, and consistency verification. The pilot annotation stage was conducted on 10 reports, and after revising the annotation guidelines based on disagreement cases, the full-scale formal annotation was carried out. Annotation consistency was evaluated using Cohen's Kappa coefficient, with the Kappa value of the two annotators during the formal annotation stage being 0.87, reaching a high level of agreement (Kappa ≥ 0.81), indicating that the annotation guidelines possess good operability and reliability. The corpus statistics are shown in Table 1.

Table 1. Annotation corpus statistics.

Statistical Item	Value
Number of reports	50
Total sentences	4236
Total entities	5847
Number of full-name variants (proportion)	2105 (36.0%)
Number of abbreviation variants (proportion)	2631 (45.0%)
Number of code variants (proportion)	1111 (19.0%)
Cohen's Kappa	0.87
Training set/Validation set/Test set	3389/424/423

Cohen's Kappa was calculated at the character level, comparing the BIO tags assigned by each annotator on a character-by-character basis across the entire annotated text. The coefficient was computed before adjudication—that is, on the raw independent annotations prior to consensus resolution—thereby reflecting the genuine inter-annotator agreement and providing a conservative estimate of annotation reliability.

4. Models and Experiments

4.1. Model Setup

This study selected conditional random fields (CRF++) as the domain NER model. CRF models the conditional probability of a given observation sequence and label sequence, effectively capturing dependencies between labels; NER can achieve good recognition performance only when feature engineering is sufficient. The reason for choosing CRF over deep learning models is that this study focuses on verifying the effectiveness of the corpus, and CRF, as a lightweight statistical model, can be trained without a GPU; the entire process can proceed smoothly by simply setting up a complete experimental workflow.

The CRF feature template includes four types of features: 1) character-level features, including the current character and two characters before and after it; 2) part-of-speech features, extracting the POS tag of the current word based on word segmentation results; 3) contextual window features, covering combination features of two characters before and after; 4) domain dictionary features, constructing binary features based on the full names, abbreviations, and code tables of formation names, indicating whether the current character belongs to a known formation name. Additionally, a rule-based baseline model was set as the comparative lower bound. The rule-based baseline adopts a combination of regular expressions and formation name dictionary matching, implemented through Python scripts without requiring additional software dependencies. The formation-name dictionary used for domain dictionary features was constructed exclusively from the training set annotations. No entities from the validation or test sets were included in the dictionary, and no external resources (*e.g.*, published stratigraphic lexicons or databases) were used during dictionary construction. This design ensures that the dictionary features reflect only the knowledge available during training, ruling out evaluation leakage.

4.2. Experimental Setup

The evaluation metrics adopted entity-level precision (P), recall (R), and F1 score. Precision is defined as the ratio of correctly identified entities to the total number of entities output by the model, recall is defined as the ratio of correctly identified entities to the total number of annotated entities, and the F1 score is the harmonic mean of the two. The experimental environment was CRF++ 0.58, running on a standard CPU without GPU acceleration. The test set was used for final performance evaluation, and the validation set was used for feature template tuning.

The training, validation, and test sets were split at the report level rather than at the sentence level. Specifically, all sentences from a given completion report or WOO chapter were assigned to the same subset. This report-level split prevents similar wording patterns and repeated formation names from appearing across different sets, thereby avoiding data leakage and providing a more realistic evaluation of the model's generalization ability to unseen reports. The recognition performance of each model is shown in Table 2. The rule-based baseline model achieved an F1 score of 74.8%, with its precision (82.3%) higher than its recall (68.5%), mainly because dictionary matching tends to exactly match full-name and abbreviation variants but has insufficient coverage for informal descriptions and nested structures. The CRF basic feature model (character-level + POS features) achieved an improved F1 score of 82.1%, an increase of 7.3 percentage points over the rule-based baseline, verifying the adaptability of statistical learning to variant diversity.

Table 2. Recognition performance comparison of each model (%).

Model	Precision (P)	Recall (R)	F1
Rule-based baseline (regex + dictionary)	82.3	68.5	74.8
CRF++ (basic feature template)	85.1	79.3	82.1
CRF++ (full features + domain dictionary)	88.6	84.7	86.6
CRF++ + rule-based post-processing	87.9	86.2	87.0

After incorporating domain dictionary features and the full contextual window, the F1 score of the CRF model further improved to 86.6%, an increase of 4.5 percentage points over the basic feature model. The recall improvement was particularly significant (from 79.3% to 84.7%), indicating that domain dictionary features effectively covered code variants and low-frequency abbreviation variants. The introduction of the rule-based post-processing module brought the F1 score to 87.0%, with recall improving to 86.2%, but precision decreased slightly (from 88.6% to 87.9%), due to a small number of false matches introduced by code regular expression matching.

The ablation experiment results are as follows: after removing domain dictionary features, the F1 score decreased by 3.8 percentage points (86.6% → 82.8%), indicating that the domain dictionary is the highest-contributing feature module; reducing the contextual window from ± 2 to ± 1 resulted in a 1.7 percentage point decrease in F1 score (86.6% → 84.9%), suggesting that a larger contextual window helps capture boundary information of nested structures; removing the rule-based post-processing module resulted in a 0.4 percentage point decrease in F1 score (87.0% → 86.6%), with this module's main contribution being the supplementation of code recall. All feature modules positively contribute to model performance.

4.3. Experimental Results

The classification and statistical results of errors made by the fully-featured CRF model on the test set are as follows: boundary truncation is the highest-proportion error type (28%), typically manifesting as truncating "the lower submember of the Sha-3 Member" to "the Sha-3 Member", omitting the hierarchical suffix. The root cause of this error lies in the complexity of nested structures—when multiple hierarchical suffixes appear consecutively, the CRF's label transition probability tends to prematurely terminate entity annotation at the higher hierarchical level. Missed recall of informal descriptions accounts for 19%, primarily involving cases where reports use colloquial or non-standard expressions to refer to formations, such as "the upper Sha-3" and "the lower Guantao Formation"; such expressions do not strictly follow standard stratigraphic naming conventions, resulting in the domain dictionary being unable to cover them. Lithology mislabeled as formation names accounts for 15%, with typical errors being the mislabeling of lithological terms such as "sandstone" and "limestone" as formation entities, primarily occurring in contexts where lithological names and formation names are morphologically similar. Unrecognized codes account for 11%, concentrated on non-standard codes with low frequency of use. The remaining errors account for 27%, including inconsistent annotation boundaries, compound entity splitting errors, and others.

5. Conclusions

This study addressed the problem of automatic recognition of formation names in completion reports, completing three tasks: corpus construction, annotation guideline design, and model comparison verification. It provides additional reference pathways for corpus annotation and linguistic analysis of geological texts. Through this corpus construction and NER model comparison, the main conclusions are as follows:

First, this study constructed the annotated corpus for formation names in completion reports, covering three types of variants: full names, abbreviations, and codes. The annotation consistency Cohen's Kappa reached 0.87, verifying the operability of the annotation guidelines and the reliability of the corpus. Second, based on AntConc frequency analysis and KH Coder co-occurrence network analysis, the verb collocation distribution and semantic association patterns of formation names were revealed, and data-driven annotation guidelines were designed accordingly. Linguistic feature analysis findings were directly mapped to CRF feature templates, forming a closed-loop methodology of "analysis-annotation-modeling." Third, the fully-featured CRF model combined with rule-based post-processing achieved an F1 score of 87.0%, an improvement of 12.2 percentage points over the rule-based baseline. The ablation experiments verified the effectiveness of domain dictionary features (contributing 3.8 percentage points) and contextual window features (contributing 1.7 percentage points). Error analysis showed that boundary truncation (28%) and missed recall of informal descriptions (19%) are the primary error types.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] OPEC (2026) World Oil Outlook 2026. OPEC.

[2] Lafferty, J., McCallum, A. and Pereira, F.C.N. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning*, San Francisco, 28 June 2001-1 July 2001 282-289.

[3] Huang, Z., Xu, W. and Yu, K. (2015) Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991.

[4] Devlin, J., Chang, M.W., Lee, K., *et al.* (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, Minneapolis, June 2 - June 7 2019, 4171-4186. https://aclanthology.org/N19-1423/

[5] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., *et al.* (2019) BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, **36**, 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

[6] Leitner, E., Rehm, G. and Moreno-Schneider, J. (2019) Fine-Grained Named Entity Recognition in Legal Documents. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H. and Sure-Vetter, Y., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 272-287. https://doi.org/10.1007/978-3-030-33220-4_20

[7] Enkhsaikhan, M., Holden, E., Duuring, P. and Liu, W. (2021) Understanding Ore-Forming Conditions Using Machine Reading of Text. *Ore Geology Reviews*, **135**, 104200. https://doi.org/10.1016/j.oregeorev.2021.104200