Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 1, 2026, 115-127.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

PWNC: A Large-Scale Persian Corpus for Joint WSD and NER Using Semi-Supervised and Supervised Learning

Arash Keshtkar, Saeedeh Sadat Sadidpour * and Hossien Shirazi

Faculty of Electrical & Computer Engineering, Malek Ashtar University of Technology, Iran.

Article Info

Article History:
Received 05 April 2025
Revised 01 June 2025
Accepted 02 November 2025

DOI:10.22044/jadm.2025.15873.2701

Keywords:
Word Sense Disambiguation; Named Entity Recognition; Semantic Disambiguation; Persian Language Processing.

**Corresponding author: sadidpour@mut.ac.ir (S. Sadat Sadipour).*

Abstract

Word Sense Disambiguation (WSD) is a longstanding challenge in natural language processing, particularly in morphologically rich and low-resource languages such as Persian. The inherent ambiguity of Persian named entities exacerbated by domain-specific contexts and limited labeled data complicates both semantic interpretation and information extraction. In this study, we introduce the PWNC corpus, a large-scale, integrated dataset designed for both Named Entity Recognition (NER) and WSD in Persian. The corpus was automatically constructed through a semi-supervised framework, incorporating contextual similarity measures and clustering algorithms to annotate ambiguous entities across ten semantic categories. Utilizing a semi-supervised framework, the proposed homograph semantic categorization method achieved robust performance, with a precision of 83%, recall of 81%, and an F1-score of 82% across over 305K annotated paragraphs. Detailed error analysis revealed challenges in disambiguating closely related senses and weak entities, which were mitigated through contextual embedding strategies. This work provides the first publicly available dual-task corpus for Persian NER and WSD, offering a scalable solution for disambiguation in low-resource tasks and laying the baseline for future research in Persian semantic processing.

1. Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that involves identifying and classifying named entities in text into predefined categories such as persons, organizations, locations, and events. Over the past few decades, NER has received extensive research attention, leading to a range of methodological advances. Early systems relied on handcrafted linguistic rules and domain-specific features, requiring significant manual effort in feature engineering and limiting generalizability across domains.

The advent of deep learning architectures since 2011 has transformed the NER landscape, enabling end-to-end models that automatically learn contextual representations from data. These approaches particularly those based on recurrent networks and, more recently, transformers have become the dominant paradigm due to their superior performance and reduced reliance on manual feature design [1].

Despite these advances, a persistent challenge in NER is entity ambiguity: many surface forms correspond to multiple real-world entities spanning different semantic categories. For example, the same word may denote a location in one context and an organization in another. Resolving such ambiguities requires Named Entity Disambiguation (NED), a process that leverages contextual cues to assign the correct interpretation to each entity mention.

In Persian, entity ambiguity is further complicated by linguistic phenomena such as homography and metonymy. Consider the word “Tehran”: in most contexts, it denotes the capital city of Iran (a location), but it can also refer metonymically to the provincial government (an organization). For instance, consider the sentence: “Alireza Fakhari, the Governor of Tehran, attended the event.” Here, the word “Tehran” does not refer to the city as a mere geographic location but to Tehran Province, a geopolitical entity whose administrative head is the Governor. This reflects a common pattern in Persian, where location names are used metonymically to denote governmental bodies (e.g., “Tehran announced a policy” = the provincial government). Accurately recognizing this requires not only labeling “Tehran” as a location (LOC) in NER (since it denotes a place-based institution) but also disambiguating its institutional sense in WSD a capability absent in existing Persian NER or WSD resources, which treat these tasks in isolation. To illustrate this challenge in a real-world context, Figure 1 presents an annotated excerpt from a Persian news report. The example highlights how joint NER–WSD modeling is essential for correct semantic interpretation Figure 1.

“*The National Guild Day was held on the morning of Saturday, July 24, 2023, with the presence of Mohammad Bagher Ghalibaf, the Speaker of the Islamic Consultative Assembly; Alireza Fakhari, the Governor of Tehran; Gholamali Haddad Adel, the President of the Academy of Persian Language and Literature; and Ghasem Noudi Farahani, the representative of the Minister of Industry, Mine and Trade, at the conference hall of the Islamic Republic of Iran Broadcasting (IRIB)*”

"روز ملی اصناف صبح روز شنبه ۳ تیر ماه ۱۴۰۲ با حضور محمد باقر قالیباف رئیس مجلس شورای اسلامی، علیرضا فخاری استاندار تهران، غلامعلی حداد عادل رئیس فرهنگستان ادب فارسی و قاسم نوده فراهانی نماینده وزیر صمت در سالن همایش های صداوسیما برگزار شد"

1. NER: In this text, the word "Tehran" is identified as a location (LOC).
2. NED: In this stage, disambiguation must take place, as the term "Tehran" has multiple meanings:
   - Tehran: The capital of Iran
   - Tehran: A street in South Korea
   - Tehran: A symbol representing the Iranian government

Figure 1. Example of contextual ambiguity in Persian NER and WSD.

A major challenge in Persian NLP is domain specificity. Most pre-trained language models (e.g., ParsBERT) are trained on general-domain corpora such as news and Wikipedia. While they perform well on in-domain tasks, their contextual representations degrade significantly when applied to specialized domains like medicine, law, or finance precisely where accurate NER and WSD are most critical.

Saeidi et al. [2] emphasize that domain-specific pretraining is essential for medical NLP, given the rapid growth of clinical texts requiring automated analysis. Models like Word2Vec, ELMo, and BERT must be adapted to medical corpora to capture domain-specific semantics.

However, because most benchmarks use general-domain data (e.g., Wikipedia, Common Crawl), evaluating model performance on specialized texts remains challenging [2].

Effective Word Sense Disambiguation (WSD) relies on rich lexical-semantic knowledge. Structured resources such as WordNet [3], BabelNet [4], and Wikipedia are commonly used to ground word senses in real-world concepts, enabling systems to select the correct meaning based on contextual compatibility [5].

Recent advances leverage pre-trained language models to generate contextual embeddings that implicitly encode semantic distinctions, offering a promising path for low-resource languages like Persian. These models can facilitate the creation of specialized datasets in domains such as economics, politics, and law, even when labeled data is scarce.

To address these gaps, we propose the Persian-WSD-NER-Corpus (PWNC), a large-scale dataset spanning diverse domains including economics, politics, sports, and culture that supports joint NER and WSD annotation. Unlike prior Persian corpora, which are limited to a single task or domain, PWNC is the first resource to unify entity recognition and sense disambiguation within a single framework. This integrated design enables the development of models that simultaneously identify entities and resolve their meanings, thereby enhancing performance on downstream NLP applications.

The remainder of this paper is structured as follows. Section 2 presents a comprehensive review of existing corpora for NER and WSD in Persian and closely related languages, with a focus on annotation schemes, domain coverage, and methodological approaches to resource construction. Section 3 details the methodology underlying the creation of the Persian-WSD-NER-Corpus (PWNC), including data collection, preprocessing pipelines, knowledge base integration, homograph identification, and the dual-task annotation protocol. In Section 4, we describe our experimental setup, present quantitative and qualitative evaluation results, and provide an in-depth error analysis to elucidate model strengths and limitations. Section 5 discusses the broader implications of our findings, including the impact of joint NER–WSD modeling on Persian NLP and the potential for cross-domain generalization. Finally, Section 6 concludes the paper and outlines promising avenues for future work, such as extension to other low-resource languages.

2. Literature review

Recent advances in deep learning have substantially improved WSD, particularly through neural architectures that model contextual semantics. Early neural approaches framed WSD as a multilabel classification task, where each sense of a polysemous word is treated as a distinct label, and contextual embeddings are used to predict the correct sense [6]. More recently, transformer-based models such as BERT have enabled context-aware representations that achieve state-of-the-art performance on WSD benchmarks for high-resource languages. However, such progress has not yet been fully realized for Persian due to the scarcity of annotated WSD resources.

In contrast to high-resource languages where large-scale NER datasets have spurred the development of robust models, Persian NER research has been constrained by limited annotated corpora. To date, only a few Persian-specific NER resources exist (e.g., Arman [7], Peyma [8]), and the corresponding models are listed in Table 1. Critically, none of these datasets explicitly address the interaction between NER and WSD, despite the fact that homographic entities (e.g., “Apple” as a company vs. a fruit) pose a known challenge in low-resource settings Table 1.

Table 1. Persian Approaches in the Task of NER.

Approach	Description
Statistical[9]	Which follows a statistical methodology and utilizes labeled datasets
Machine Learning[10]	Which is based on machine learning and human knowledge
Deep Learning[11]	Which has evolved in recent years with the advancement of deep neural network models

Over the past three decades, extensive research has been conducted on high-resource languages such as English, Chinese, and French in the field of NER. In contrast, studies on low-resource languages like Persian, despite their linguistic and societal significance, have received comparatively less attention, particularly in multi-task NLP applications. Some of the key studies in this area are outlined below:

Mortazavi and ShamsFard [12] introduced the first system for the detection and classification of named entities in the Persian language, employing rule-based methods. Subsequently, with the emergence of machine learning techniques, researchers began developing data-driven NER systems that offered greater flexibility and scalability compared to handcrafted rules.

Poostchi et al. [13] introduced the Arman dataset, a new NER resource for the Persian language. Around the same time, the advent of deep learning models, particularly BERT, which leverages attention mechanisms and transformer architectures, sparked significant interest in applying pre-trained language models to low-resource NER tasks, including Persian.

Farahani et al. [14] introduced ParsBERT, a BERT-based model trained specifically on Persian corpora. They fine-tuned ParsBERT for the NER task and evaluated its performance on the Arman [7] and Peyma [8] datasets, achieving state-of-the-art results.

A major challenge in Persian NER stems from the prevalence of homograph words that share the same spelling but have different meanings depending on context. Since the correct interpretation of such words is essential for accurate entity recognition, disambiguating homographs becomes a critical subtask. This problem is formally framed as a WSD task [9]. However, due to the scarcity of freely available digital resources for Persian, the development of WSD datasets and dedicated disambiguation methods has remained limited. Notable prior efforts in this area include the following:

Maki and Homayounpour [9] proposed an unsupervised hybrid method to construct a small-scale WSD dataset comprising 15 homographs, where word senses were extracted from a Persian thesaurus. They also used this approach to build a context-rich corpus from domain-specific texts.

Masoudi and Zandokili [10] proposed an unsupervised method for disambiguating Persian homographs by leveraging two contextual features: (i) the surrounding words and punctuation marks adjacent to the target word, and (ii) the overall topic of the text. They employed Latent Dirichlet Allocation (LDA) to infer document-level topics and used these features to construct a homograph disambiguation dataset.

Mahmoudvand and Hoorali [15] developed an unsupervised approach to build a dataset for three Persian homographs, collecting 5,368 web-crawled documents. Their method used a two-stage hybrid feature extraction strategy: first, it captured local syntactic context by extracting words within a fixed window around the target word and encoding their positional information; second, it constructed binary vector features indicating the presence or absence of words associated with each possible sense of the homograph.

Moradi et al. [16] explored cross-lingual transfer for WSD by applying a pre-trained English word2vec model in conjunction with a Persian–English dictionary to disambiguate four Persian homographs. Their approach was grounded in the assumption that words in a sentence are semantically and syntactically interdependent; thus, the meaning of ambiguous words can be inferred from the contextual representations of their unambiguous neighbors.

Ghaeyomi [17] introduced a Word Sense Induction (WSI) method for 20 Persian homographs, compiling a dataset of 2,000 sentences (100 per homograph). The dataset was manually annotated following the annotation guidelines of SemEval 2010, ensuring consistency with established WSD evaluation standards.

Roohizadeh and Shamsfard [18] developed a WSD system for Persian that incorporated part-of-speech (POS) tagging as a preprocessing step to enrich contextual representations. They constructed a dataset from news articles, performed semantic disambiguation using sense inventories from FarsNet (the Persian WordNet), and trained their model on the PerSemCor corpus [11], a manually annotated Persian semantic resource.

This paper introduces the Persian-WSD-NER-Corpus (PWNC), a new resource designed to support joint NER and WSD in Persian. The corpus comprises two components: (i) a large-scale training set generated through an unsupervised distant supervision method, and (ii) a high-quality evaluation set manually annotated by linguistic experts. By providing rich contextual annotations, PWNC aims to enhance the training of hybrid deep learning models that leverage contextual representations, thereby reducing reliance on costly, fully supervised datasets.

A persistent challenge in Persian NLP is the scarcity of semantically annotated data, especially in specialized domains such as medicine, law, and cyberspace. To address this, our approach exploits the transfer learning capabilities of large pre-trained language models, particularly ParsBERT, which have demonstrated strong cross-domain generalization. This enables the efficient creation of task-specific annotations with minimal manual effort, even in low-resource settings.

PWNC integrates NER and WSD annotations within a unified framework and contains over 300,000 annotated instances, making it the largest publicly available Persian corpus that supports both tasks simultaneously. To the best of our knowledge, no existing Persian resource combines fine-grained entity typing with word sense labels at this scale.

Table 2 compares PWNC against existing Persian resources including ArmanPersoNER, Peyma, SBU-WSD-Corpus, and FarsNet across key dimensions such as corpus size, supported tasks, and NER tag types. The NER categories include PER (person), ORG (organization), FAC (facility), EVE (event), and LOC (location), following the OntoNotes annotation scheme. As the table shows, PWNC is not only substantially larger but also the only dataset that jointly supports both NER and WSD, addressing a critical gap in Persian NLP infrastructure.

Table 2. Comparison of Persian NER/WSD datasets.

Dataset	Size	Labels	Domains	Annotation Method
ArmanPersoNER[7]	7K+ sentences	PER, ORG, LOC, FAC, EVE, PRO	BijanKhan	Manual
Peyma[8]	7K+ sentences	PER, ORG, LOC, TIM, DAT, MON, PCT	News	Manual
FarsNet[19]	10K	Synset	Wordnet	Semi-automatic
ParsNER-Social[20]	20K+ sentences	PER, ORG, LOC, MISC	Social media (Telegram channels)	Manual
SBU-WSD-Corpus[18]	19 docs	POS relation to Farsnet	News	Manual
Our	**305K** sentences	PER, ORG, LOC, FAC, EVE, PRO, TIM, DAT, MON, PCT **Word Sense**	News (diverse)	**Semi-supecrvised**

3. Methodology

The proposed approach consists of independent components that purposefully collaborate to address the challenges of WSD in Persian. In this context, the process begins with preparing a suitable corpus, followed by preprocessing the texts to enhance the quality of the training data for deep learning methods. Using a contextual knowledge extraction method, semantic labels are predicted for each training document, and the features of the text for each target word are identified.

A pre-trained deep learning model, optimized for NER, is employed to predict annotations for word sequences within each sentence, leveraging the initial dataset. Subsequently, this model is integrated into a semi-supervised learning framework. Texts containing homographs with varying labels are then selected and processed in the final stage to assign definitive semantic labels to each sentence. The comprehensive structure of the proposed approach's components is depicted in Figure 2.

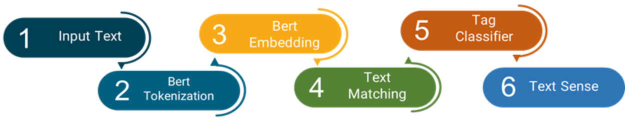


Figure 2. Overview of the Proposed Approach Structure.

Although various datasets in Persian, such as Miras [20], Hamshahri [21], and Persian Wikipedia, have diverse applications and significantly contribute to the development of natural language processing, knowledge extraction, and question-answering research, these collections are not specifically suitable for word sense disambiguation. In the domain of word sense disambiguation, a dataset with a sufficient number of samples is required to describe the meaning of the target word concerning the context of the text, enabling deep learning models to generalize effectively for this task. Moreover, long documents are not suitable for such tasks. Existing methods prior to deep learning only utilized the target word and a limited number of surrounding words, leaving a vast amount of irrelevant information unused.

To address the issue of insufficient training data for WSD in Persian, this study employs a web crawling method to construct the necessary corpus. This involves reading all publicly accessible web pages and collecting their Persian texts. Consequently, datasets such as Miras, Persian news agencies, e-commerce websites, blogs, and others can be utilized. However, this volume of Persian data is often unsuitable for deep natural language processing models, and collecting large amounts of data can be costly.

Therefore, the best approach is to use publicly available multilingual crawled datasets like mC4, an extended version of the C4 pre-training datasets covering 101 languages and making adjustments to enhance compatibility with multilingual texts [23]. Recently, the introduction of large language models by Google and OpenAI has led to significant advancements in NLP tasks [24], with deep learning models based on transformer architectures such as BERT being developed [25]. BERT can be trained for specific tasks such as classification, question answering, and sequence tagging. Following the release of BERT, various models based on its architecture have been developed to improve performance or address specific tasks. Therefore, in this study, to tackle the issue of the high cost of collecting large datasets and training models, we utilized the multi-task pre-trained model ParsBERT [14] for knowledge extraction from texts.

ParsBERT is introduced as a monolingual model designed explicitly for Persian and has been trained on a diverse set of Persian language datasets. The model was also fine-tuned for named entity recognition. Thus, a suitable dataset for training a word sense disambiguation system and named entity recognition in Persian has been created with minimal time investment. The documents in this dataset consist of phrases and sentences that include each target word in the Persian language.

3.1. Text Preprocessing

Considering that data collection in this study was conducted through web crawling and in an unsupervised manner, it is essential to review, revise, and in some cases correct the created corpus to ensure the quality of the collected data. This review may involve correcting characters, punctuation, spacing, and removing extraneous phrases, as additional information in the documents can cause issues during the knowledge extraction phase.

Crawled datasets often face the significant issue of a lack of a unified structure within the text. This problem arises due to differences in the structure and layout of pages where texts are presented in hypertext markup language (HTML). Additionally, Cascading Style Sheets (CSS) apply desired layouts and styles to the texts. This variability leads to differences in the arrangement of the main text on each extracted page, which can complicate and slow down the preprocessing process. Moreover, removing extraneous information from pages, such as web page tags and unnecessary texts in various sections, is crucial; thus, this step is included as part of the preprocessing in this study.

In this study, Persian is considered the reference language for word sense disambiguation. This choice presents more challenges compared to other languages, such as Latin languages. One of these challenges is the writing style of Persian. In Persian, some words and letters are connected, which complicates the identification of the target word in Persian texts. For instance, when a word is plural, certain letters attach to the target word, adhering to its root. This issue makes it difficult to determine the correct position of the target word, thereby limiting the use of sequential features. To address this problem, word segmentation of target words has been employed in this research.

For this purpose, during the extraction of texts, each target word in the sentence is processed to remove all attached extra words, letters, and symbols. Therefore, to ensure proper processing and understanding of the texts by language models, preprocessing is essential. In this regard, commonly used algorithms and libraries for Persian preprocessing, such as ParsiNorm, have been utilized.

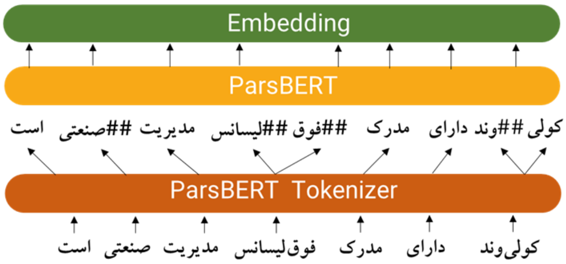

Embedding
ParsBERT
کولی ##وند دارای مدرک ##فوق ##لیسانس مدیریت ##صنعتی است
ParsBERT Tokenizer
کولی‌وند دارای مدرک فوق‌لیسانس مدیریت صنعتی است

Figure 3. ParsBERT Embedding Extraction Process.

The preprocessing steps include removing various characters and symbols, such as links and email addresses (@, #, $, and %), as well as stop words like "از". Additionally, to systematically separate and segment words, especially compound words and plural pronouns, standard spacing and character substitution guidelines have been employed. Furthermore, stemming has been omitted due to the nature and objectives of this research. Consequently, all target words in the text are presented without prefixes or suffixes. This method enhances the accuracy of identifying target words in the text and increases the reliability of the results.

3.2. Knowledge Extraction

The knowledge extraction component of the proposed approach begins with the user inputting information into the system, initiating the process of labeling training documents and extracting features from the target word and the input text. Once the preprocessing stage is complete, the data is ready for use in creating the training dataset for word sense disambiguation. However, before this, it is necessary to extract relevant features from the input data. Therefore, feature extraction from the texts is performed using the pre-trained ParsBERT model.

As shown in Figure 3, the input text is tagged by the ParsBERT model, and then the embedding of the input text is extracted from the model. In other words, because this model has been trained on large and diverse datasets, it is enriched with linguistic knowledge, enabling context-aware knowledge extraction. Not only is the extracted knowledge applicable to the current task, but it can also be utilized in other natural language processing tasks. Recently, this embedding extraction method has demonstrated superior adaptability to specialized texts in various fields compared to other methods such as n-grams, TF-IDF, and others, due to its contextual approach. Another advantage is the ability to fine-tune this model in an unsupervised manner, which other methods cannot achieve Figure 3.

3.3. Dataset Overview

The PWNC consists of approximately 305,000 paragraphs (about 290,000 original and 10,000 revised) extracted from over 462,000 articles from Persian news agencies such as Fars, Tabnak, and Hamshahri. This corpus was developed to support both NER and WSD tasks, addressing the scarcity of comprehensive Persian datasets.

3.3.1. Collection

The corpus was assembled by crawling publicly available Persian news articles using Python's BeautifulSoup library. To ensure domain diversity, sources with distinct editorial orientations were selected: Fars News Agency (politics and current affairs), Hamshahri Online (culture, society, and urban life), and Tabnak (economics, business, and general news). From this collection, stratified sampling was applied to achieve balanced representation across five key domains: politics, economics, culture, sports, and religion.

Table 3. NER annotates description.

Class	Title	Description	Example
PER	Person	Names of individuals, which include full names	Mohammad Yousefi, Seyyed Kamal Mousavi
ORG	Organization	Names of organizations, ministries, departments and companies	Social Security Organization, Ministry of Agriculture, Tamin Niro Company
LOC	Location	Names of locations	Country, city, village, etc. (Tehran, South Korea)
EVE	Event	Names of specific events	Dahe Fajr, the blessed birth of ...., parliamentary elections
FAC	Facility	Names of public facilities	Azadi Stadium, Charsou Cinema
PRO	Product	Names of products or inaugurations	Karun Dam, COVID-19 vaccine, Omid satellite
MON	Money	Monetary values	Dollar, Toman, 100 euros
PCT	Percent	Percentage values	80 percent, 200%
DAT	Date	Date values	Friday, January 3, 2014; February 17, 2014
TIM	Time	Time values	10:19, 11:00 PM
O	Other	Unlabeled	Hello, became, is, etc

3.3.2. Annotation Process

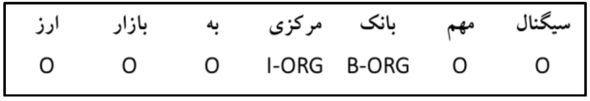
A hybrid annotation approach combined automated and human efforts. ParsBERT, fine-tuned on ArmanPersoNER, initially annotated 3% of the dataset (10K paragraphs) for NER (labels: PER, ORG, LOC, EVE) and WSD (homograph senses). Two annotators, each with over two years of expertise in Persian NLP, manually validated 3% of these annotations (10K paragraphs). The remaining approximately 97% (290K paragraphs) were annotated using a semi-supervised pipeline: K-means clustering (K=50) grouped embeddings, and cosine similarity (threshold=0.7). A manual review of 1% of the semi-supervised annotations (approximately 3K paragraphs) verified annotation consistency.

3.3.3. Annotation Quality Control

To ensure annotation accuracy, a subset of the corpus was independently annotated by two Persian NLP experts, each with over two years of experience. Inter-annotator agreement was calculated to assess consistency. Discrepancies were attributed to differing contextual interpretations (e.g., variable entity type assignments for homographs) or errors in label assignment. Given that most tokens are non-entities (labeled "O" in the BIO scheme), disagreement rates were also computed for tokens identified as named entities by either annotator, using the union of their entity annotations as the reference set, as no definitive ground truth exists for entity identification. Data integrity was maintained using the Persian Hashtag project's news document classifier to categorize documents into six domains: politics, economy, sports, culture and arts, science and technology, and society. For each domain, the mean document length and standard deviation were calculated, and documents were sampled following a normal distribution to reflect real-world length distributions. To ensure balanced domain representation, documents were selected over time to match the overall domain distribution of the source data. To prevent over-reliance on dominant news agencies (e.g., Fars), each source was limited to 25% of the final dataset, with sampling proportional to each source's contribution per domain. Boilerplate content, such as advertisements and navigation menus, was removed during preprocessing using BeautifulSoup and ParsiNorm. Iterative reviews corrected errors in text segmentation, punctuation, and character encoding, particularly for Persian's connected writing system.

3.3.4. Annotation Scheme

We employ ParsBERT, a Persian-adapted BERT model, as the backbone of our NER pipeline. Input texts are tokenized and fed into ParsBERT to generate contextual embeddings, which are then passed through a token-level classification layer fine-tuned for named entity recognition. The model assigns each token a label from a predefined set of 11 entity categories (see Table 3 for the full inventory and definitions), using the BIO (Begin–Inside–Outside) tagging scheme to consistently represent multi-token entities. To illustrate this labeling convention, Figure 4 presents a sample sentence from our corpus with its corresponding BIO annotations. For instance, the multi-token entity «بانک مرکزی» ("Central Bank") is labeled as B-ORG for the first token «بانک» and I-ORG for the subsequent token «مرکزی», while non-entity tokens such as «سیگنال» ("Signal") are marked as O. This scheme enables precise boundary detection for complex Persian named entities, including compound nouns and institutional phrases.



ارز	بازار	به	مرکزی	بانک	مهم	سیگنال
O	O	O	I-ORG	B-ORG	O	O

Figure 4. Example of BIO tagging in Persian NER.

3.4. Semantic Categorization

In this approach, a semi-supervised learning method is employed to address the shortage of training data in the Persian language. The semantic texts of the words have been classified using two methods:

1. **Clustering**: Using a set of clustering techniques, including KMeans, Agglomerative and Spectral Clustering, these methods have been individually applied to all extracted embeddings. The goal of executing clustering on the data is to separate the data into expected groups unsupervisedly.
2. **Textual Similarity**: Similarity assessment has been conducted using embeddings extracted from the corpus texts and revised texts through cosine and Jaccard methods. After clustering and assessing text similarity, the results were evaluated using a labeled dataset created by several experts, as shown in the following equations, where C is cosine similarity (1), J jaccard similarity (2), A represents vector of labeled sentence i and B represents vector of predicted sentence j by ParsBERT.

Table 4. Classification of homograph words.

Target Word	Class	Text
Thamen al-Hojjaj ثامن الحجج	EVE	He congratulated on the anniversary of the birth of Thamen al-Hojjaj, Imam Reza (AS) ایشان با تبریک سالروز میلاد ثامن الحجج، حضرت امام رضا (ع)
	PER	The celestial shrine of Thamen al-Hojjaj was adorned with the rosewater of Kashan بارگاه ملکوتی ثامن الحجج، معطر به گلاب کاشان شد
	ORG	Organization of Thamen al-Hojjaj Credit Cooperative ساماندهی تعاونی اعتبار ثامن الحجج
	PRO	Inauguration of the Thamen al-Hojjaj Medical Center افتتاح مرکز درمانی ثامن الحجج
	LOC	Travel plans include visiting the shrine of Thamen al-Hojjaj and meeting with the custodian of the Astan Quds برنامه‌های سفر، بازدید از بارگاه ثامن الحجج، دیدار با تولیت آستان قدس
Islamic Republic جمهوری اسلامی	EVE	The celebration of the victory of the Islamic Republic Revolution received extensive coverage in Western media جنش پیروزی انقلاب جمهوری اسلامی در رسانه‌های غربی بازتاب فراوان داشت
	ORG	The Islamic Republic aims to collaborate with Japan on waste-to-energy conversion جمهوری اسلامی تمایل دارد در مورد تبدیل زباله به انرژی، با کشور ژاپن همکاری داشته باشد
	FAC	The CEO of the Isfahan Bus Company emphsized that this service is free of charge. He noted that this number of urban buses operates from 20 central points in the city, including the Shahid, Ahmadabad, Khorasgan, Islamic Republic, Azadi squares and the streets of Imam Khomeini, Kashavarz Boulevard and Malek Shahr, as well as all routes of the city’s rapid bus transit مدیرعامل شرکت اتوبوسرانی اصفهان تأکید کرد که این خدمت رایگان است. او اشاره کرد که این تعداد اتوبوس‌های شهری از ۲۰ نقطه مرکزی در شهر فعالیت می‌کنند، از جمله میدان‌های شهدا، احمدآباد، خوراسگان، جمهوری اسلامی، آزادی و خیابان‌های امام خمینی، کشاورز و ملک شهر، همچنین تمامی مسیرهای حمل و نقل بی آر تی شهری
	LOC	Refrigerated warehouse on Islamic Republic Boulevard behind the car dealership انبار یخچالی در بلوار جمهوری اسلامی پشت بنگاه اتومبیل

$$\mathrm{C}(A_i,B_j)=\frac{A_i.B_j}{\|A_i\|\times\|B_j\|} \quad (1)$$

$$J(A_i,B_j)=\frac{|A_i\bigcap B_j|}{|A_i\bigcup B_j|}=\frac{|A_i\bigcap B_j|}{|A_i|+|B_j|-|A_i\bigcap B_j|} \quad (2)$$

Therefore, a contextual matching approach has been used to classify the collected data. This method categorizes ambiguous named entity data contextually into 11 groups.

3.5. Homograph Disambiguation

This subsection describes the methodology for developing the word sense disambiguation (WSD) component of the corpus and selecting the appropriate sense for homograph words using a threshold-based approach to contextual similarity.

3.5.1 Identification

Homographs with multiple senses (e.g., “Tehran” as LOC or ORG, “Islamic Republic” as ORG or EVE) were identified by analyzing the corpus for words appearing in varied contexts with distinct entity labels or meanings. An initial list was compiled using a Persian thesaurus and validated against corpus contexts to ensure relevance to NER and WSD tasks.

3.5.2 WSD Development

For each homograph, all paragraphs containing the target word were extracted. ParsBERT generated contextual embeddings for these paragraphs. Sense prototypes (e.g., embeddings for “Tehran” as a location versus an organization) were derived from a manually annotated subset of 10,000 paragraphs. Duplicate sentences were removed to eliminate redundancy.

3.5.3 Candidate Homograph Selection

We begin by identifying Persian homographs that exhibit semantic or named entity ambiguity, i.e., words that can be assigned to multiple distinct categories (e.g., “Tehran” as a location or a metonymic organization). Candidate words are selected based on two criteria: (1) occurrence frequency in our source corpus and (2) verified polysemy through consultation with FarsNet and manual review by native-speaking linguists. For each selected homograph, we extract all sentence-level contexts containing the word. These instances are structured as triples: (target word, semantic class, context sentence), where semantic class denotes the disambiguated label (e.g., LOC, ORG, or a WSD sense ID). The resulting dataset is summarized in Table 4.

4. Evaluation Framework

4.1. Datasets and Experimental Settings

In our research, due to the lack of a combined Persian dataset for named entity recognition and disambiguation, 10,000 unique data points were reviewed by an expert. Therefore, the contextual matching method has been used to classify the collected data.

This approach categorizes named entity data contextually into 11 groups. Ambiguous words play a crucial role in WSD issues. Ambiguous words exist in all living natural languages around the world.

While there may be common words among these ambiguous terms in some languages, the language under study in this research is Persian then the data were evaluated based on precision (3), recall (4) and F1-score (5) metrics.

$$P = \frac{TP}{TP + FP} \qquad (3)$$

$$R = \frac{TP}{TP + FN} \qquad (4)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \qquad (5)$$

4.1. Clustering and Similarity-Based Disambiguation

In this approach, a semi-supervised learning method is utilized to tackle the shortage of training data in the Persian language. The semantic texts of the words have been classified using two methods:

1. **Clustering:** Using a set of clustering techniques, including K-Means, Agglomerative and Spectral Clustering, these methods have been individually applied to all extracted embeddings. The goal of executing clustering on the data is to unsupervisedly separate the data into expected groups.

2. **Textual Similarity:** Similarity assessment using embeddings extracted from the corpus texts and revised texts through Cosine and Jaccard methods. After clustering and evaluating text similarity, the results were assessed using an annotated dataset created by several experts.

4.2. Statistics of Ambiguous Entities

Table 5 presents the distribution of ambiguous entity mentions in our corpus, categorized by form and ambiguity type. Specifically, we identify 84 single-word and 7 multi-word entity mentions that exhibit semantic ambiguity. These ambiguous instances fall into six distinct ambiguity categories, such as location vs. organization (e.g., “Tehran”) and person vs. product (e.g., “Nokia”), reflecting common cross-category confusions in Persian NER and WSD tasks. This analysis highlights the diversity of disambiguation challenges addressed by our dataset

Table 5. Distribution of ambiguous entities.

Number of ambiguities per entity class						
Categories		2	3	4	5	6
	Single	5203	1782	641	224	84
Word Parts	Multi	7621	602	81	14	7
	Total	12824	2384	722	238	91

4.3. Implementation

The operating environment of this experiment was Ubuntu 22.04, Intel i9-12900K CPU @ 5.2 GHz processor with a graphics card of GeForce RTX 4090. The Python programming language was used and the inferencing platform is Pytorch 2.0 deep learning framework. The inferencing parameter settings of the experiment are shown in Table 6.

Table 6. Implementation of sequence embedding.

Parameters	Value
sequence_length	128
max_position_embeddings	512
batch_size	64
vocab_size	42000
hidden_size	768

4.4. Experimental Results and Error Analysis

As shown in Table 8, the table summarizes the comparative performance of clustering and knowledge-based approaches for word sense disambiguation. As the results indicate, clustering methods such as K-Means (F1-score: 0.51), Agglomerative Clustering (0.36), and Spectral Clustering (0.40) underperformed significantly in distinguishing contextual meanings. This performance gap reflects the difficulty these methods face in handling overlapping semantic representations, particularly in morphologically rich languages like Persian. In contrast, knowledge-based methods demonstrated notably stronger results. Cosine similarity achieved an F1-score of 0.82, and Jaccard similarity followed with 0.74, suggesting that comparisons of contextual embeddings are more reliable for Persian disambiguation than unsupervised clustering alone Table 7.

Table 8. The performance of semantic ambiguity detection using a clustering approaches.

Text	Target Word	Annotate	Approach				
			K-Means	Agglomerative	Spectral	Cosine	Jaccard
On the eve of the ninth parliamentary elections, we also witnessed the presence of the Rahpouyan and Isargaran groups در آستانه انتخابات مجلس نهم هم شاهد حضور جمعیت رهپویان و جمعیت ایثارگران	Ninth parliamentary elections مجلس نهم	EVE	EVE	ORG	LOC	EVE	ORG
At the beginning of the match between Iran and Hong Kong, there was intense and close competition در ابتدا بازی میان دو تیم ایران و هنگ کنگ رقابت فشرده و تنگاتنگی وجود داشت	Hong Kong هنگ کنگ	ORG	ORG	ORG	LOC	ORG	LOC
The CEO of BlackBerry criticized Apple regarding the San Bernardino terrorist iPhone case رئیس شرکت بلک بری از اپل در مورد پرونده آیفون تروریست سن برناردینو انتقاد	San Bernardino سن برناردینو	PER	LOC	PER	LOC	PER	PER

Table 7. Performance comparison of different methods.

Methods		Precision	Recall	F1
Clustering	K-Means	0.54	0.48	0.51
	Agglomerative	0.37	0.35	0.36
	Spectral	0.41	0.39	0.40
Textual similarity	Cosine	0.83	0.81	0.82
	Jaccard	0.76	0.73	0.74

4.4.1 Experimental Results

Our evaluation demonstrates that, while the joint NER–WSD model achieves strong overall performance, certain entity types remain challenging. As shown in Table 4, domain-specific and low-frequency entities such as «ثامن الحجج» (a religious organization) are frequently misclassified in the NER task, often labeled as location (LOC) instead of organization (ORG) due to their sparse representation in the training data. Similarly, in the WSD task, highly polysemous phrases like «جمهوری‌اسلامی» ("Islamic Republic") exhibit significant ambiguity: their contextual embeddings show substantial overlap across multiple semantic categories, including organization (ORG), location (LOC), and event (EVE), leading to inconsistent sense predictions. These results highlight the limitations of current contextual representations in capturing fine-grained semantic distinctions for culturally and linguistically complex Persian terms.

4.4.2 Error Analysis

A deeper analysis of failure cases reveals two interrelated sources of error. First, unsupervised clustering methods, particularly K-Means, struggle to model the semantic structure of Persian homographs. K-Means assumes spherical, isotropic clusters, which poorly reflect the non-linear and context-dependent relationships among word senses in Persian. Second, the absence of large-scale, sense-annotated corpora for Persian prevents the model from learning well-separated embedding subspaces for distinct meanings. Consequently, even advanced clustering techniques (e.g., Agglomerative and Spectral Clustering) produce overlapping or fragmented sense clusters.

These findings strongly suggest that purely unsupervised approaches are insufficient for Persian WSD. Instead, effective disambiguation in low-resource settings requires hybrid strategies that combine contextual similarity metrics with limited supervised signals, such as expert-annotated seed examples or fine-tuned language models, to guide sense separation and improve robustness.

5. Discussion

The development of the PWNC corpus marks a substantial advancement in Persian natural language processing, particularly in addressing the dual challenge of WSD and NER in a low-resource corpus. By leveraging large-scale web crawling, this study effectively mitigates the pervasive data scarcity in Persian NLP, producing a domain-diverse corpus encompassing political, economic, cultural, and general news articles. This diversity supports broader generalization and aligns with prior research emphasizing the importance of multi-domain datasets in improving model robustness.

A key innovation of this study lies in the integration of supervised and semi-supervised learning strategies. The semi-supervised framework enables efficient utilization of both labeled and unlabeled data, an especially impactful approach given the limited availability of annotated corpora in Persian. Within this framework, clustering algorithms such as K-Means, Agglomerative, and Spectral Clustering were used to explore inherent semantic structures. However, the suboptimal clustering performance, as evidenced by an F1-score below 0.52, underscores the limitations of these models in capturing fine-grained semantic distinctions in Persian.

To address these shortcomings, we propose several enhancements. Hierarchical clustering techniques offer more flexibility in handling non-spherical data distributions, which are common in the context of polysemous Persian entities. Additionally, fine-tuning language models such as ParsBERT using sense-annotated corpora, e.g., subsets of FarsNet, could yield embeddings more attuned to lexical ambiguity. We also recommend exploring density-based clustering methods such as DBSCAN, which are better suited for complex, overlapping semantic boundaries. Moreover, the incorporation of structured semantic knowledge from resources like WordNet can further enrich the contextual understanding of word senses.

Parallel to these efforts, the study demonstrated that contextual similarity measures, particularly cosine and Jaccard similarity, significantly outperform clustering methods. These techniques proved effective in disambiguating word senses by leveraging contextual embeddings, reinforcing the centrality of context in semantic inference, especially in languages like Persian with complex morphology and syntax.

The selection of ParsBERT as the backbone model was instrumental. As a transformer-based architecture trained specifically on Persian data, ParsBERT provided strong contextual representations for both NER and WSD tasks. Fine-tuning this model within our pipeline led to reliable semantic labeling, with precision, recall, and F1-scores consistently above 80%. These results validate the synergy between large-scale monolingual language models and semi-supervised corpus construction in low-resource scenarios.

Challenges such as disambiguating homographs and adapting models to domain-specific texts remain central to the broader research agenda. Similar to findings in other morphologically rich languages, tailored solutions particularly those incorporating both contextual modeling and external semantic knowledge are crucial for accurate disambiguation. The hybrid methodology introduced in this paper not only addresses these challenges but also contributes a scalable framework for future resource development.

Looking forward, the PWNC corpus offers a foundational resource for advancing Persian NLP across a range of tasks. Its applicability extends beyond the current scope, holding potential for adaptation to other under-resourced languages and specialized domains such as medical, legal, or religious texts. By releasing the dataset as an open resource, we aim to catalyze collaborative research and elevate the visibility of Persian in the global NLP community. Beyond its technical contributions, this work reflects a broader commitment to equitable access to language technologies in linguistically diverse regions.

6. Conclusion

This study introduces the PWNC corpus, a large-scale, integrated dataset for WSD and NER in the Persian language. Constructed through an automated web crawling framework, the corpus spans a wide range of thematic domains, including politics, economics, sports, and religion, thereby enhancing its applicability to diverse NLP tasks. The proposed methodology for corpus generation is extensible and can be adapted to other specialized domains such as medicine and law.

Evaluation results based on standard metrics (precision, recall, and F1-score) affirm the effectiveness of our semi-supervised approach, which successfully balances scalability with accuracy. While supervised models demonstrate higher precision, the inclusion of semi-supervised and unsupervised components enables the construction of large datasets with minimal manual annotation, which is critical in low-resource language contexts.

The PWNC corpus is designed to serve as a foundational resource for future Persian NLP research, offering annotated data suitable for a variety of downstream applications. Its dual-task structure, covering both WSD and NER, addresses a critical gap in existing Persian corpora and provides a framework that can be replicated for other under-resourced languages.

To foster open research and encourage collaboration within the NLP community, the corpus will be made freely available to researchers. By doing so, we aim not only to empower advancements in Persian language technologies but also to contribute to the global movement toward inclusive and multilingual natural language understanding.

Acknowledgment

We thank the members of the NLP laboratory at Malek Ashtar University for their technical support.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

References

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural Language Processing (almost) from Scratch,” Mar. 2011, doi: 10.48550/arXiv.1103.0398.

[2] M. Saeidi, E. Milios, and N. Zeh, “Biomedical Word Sense Disambiguation with Contextualized Representation Learning,” in *Companion Proceedings of the Web Conference 2022*, in WWW ’22. New York, NY, USA: Association for Computing Machinery, Aug. 2022, pp. 843–848. doi: 10.1145/3487553.3524703.

[3] G. A. Miller, “WordNet: a lexical database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[4] R. Navigli and S. P. Ponzetto, “BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network,” *Artificial Intelligence*, vol. 193, pp. 217–250, Dec. 2012, doi: 10.1016/j.artint.2012.07.001.

[5] “Entity Linking meets Word Sense Disambiguation: a Unified Approach | Transactions of the Association for Computational Linguistics | MIT Press.” Accessed: Feb. 21, 2023. [Online]. Available: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00179/43316/Entity-Linking-meets-Word-Sense-Disambiguation-a

[6] A. Raganato, C. Delli Bovi, and R. Navigli, “Neural Sequence Learning Models for Word Sense Disambiguation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1156–1167. doi: 10.18653/v1/D17-1120.

[7] “BiLSTM-CRF for Persian Named-Entity Recognition ArmanPersoNERCorpus: the First Entity-Annotated Persian Dataset - ACL Anthology.” Accessed: Nov. 15, 2023. [Online]. Available: https://aclanthology.org/L18-1701/

[8] M. S. Shahshahani, M. Mohseni, A. Shakery, and H. Faili, “PEYMA: A Tagged Corpus for Persian Named Entities,” Jan. 30, 2018, *arXiv*: arXiv:1801.09936. doi: 10.48550/arXiv.1801.09936.

[9] R. Makki and M. M. Homayounpour, “Word Sense Disambiguation of Farsi Homographs Using Thesaurus and Corpus,” in *Advances in Natural Language Processing*, B. Nordström and A. Ranta, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 315–323. doi: 10.1007/978-3-540-85287-2_30.

[10] B. Masoudi and A. Zandvakili, “Persian Word Sense Disambiguation using LDA topic model,” in *International Conference on Science and Engineering*, Nov. 1394. Accessed: Nov. 16, 2023. [Online]. Available: https://civilica.com/doc/424627/

[11] H. Rouhizadeh, M. Shamsfard, M. Dehghan, and M. Rouhizadeh, “Persian SemCor: A Bag of Word Sense Annotated Corpus for the Persian Language,” in *Proceedings of the 11th Global Wordnet Conference*, University of South Africa (UNISA): Global Wordnet Association, Jan. 2021, pp. 147–156. Accessed: Apr. 28, 2023. [Online]. Available: https://aclanthology.org/2021.gwc-1.17

[12] P. S. Mortazavi,Mehrnoush Shamsfard, “Named Entity Recognition in Persian Texts,” in *15th National CSI Computer Conference*, Tehran, Iran, 2009.

[13] H. Poostchi, E. Z. Borzeshi, M. Abdous, and M. Piccardi, *PersoNER: Persian named-entity recognition*. 2016. Accessed: Nov. 16, 2023. [Online]. Available: https://opus.lib.uts.edu.au/handle/10453/80094

[14] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, “ParsBERT: Transformer-based Model for Persian Language Understanding,” *Neural Process Lett*, vol. 53, no. 6, pp. 3831–3847, Dec. 2021, doi: 10.1007/s11063-021-10528-4.

[15] M. Mahmoodvand and M. Hourali, “Semi-supervised approach for Persian word sense disambiguation,” in *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, Oct. 2017, pp. 104–110. doi: 10.1109/ICCKE.2017.8167937.

[16] B. Moradi, E. Ansari, and Z. Žabokrtský, “Unsupervised Word Sense Disambiguation Using Word Embeddings,” in *2019 25th Conference of Open Innovations Association (FRUCT)*, Nov. 2019, pp. 228–233. doi: 10.23919/FRUCT48121.2019.8981526.

[17] M. Ghayoomi, “Identifying Persian Words’ Senses Automatically by Utilizing the Word Embedding Method,” *Iranian Journal of Information Processing &amp; Management*, Jan. 2019, Accessed: Nov. 16, 2023. [Online]. Available: https://www.academia.edu/67083634/Identifying_Persian_Words_Senses_Automatically_by_Utilizing_the_Word_Embedding_Method

[18] H. Rouhizadeh, M. Shamsfard, and V. Tajalli, “SBU-WSD-Corpus: A Sense Annotated Corpus for Persian All-words Word Sense Disambiguation,” *International Journal of Web Research*, vol. 5, no. 2, pp. 77–85, Dec. 2022, doi: 10.22133/ijwr.2023.354098.1128.

[19] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Reza Gholi Famian, and S. Bagherbeigi, "Semi Automatic Development Of FarsNet: The Persian Wordnet," Jan. 2010.

[20] M. Asgari-Bidhendi, B. Janfada, O. R. Roshani Talab, and B. Minaei-Bidgoli, "ParsNER-Social: A Corpus for Named Entity Recognition in Persian Social Media Texts," *Journal of AI and Data Mining*, vol. 9, no. 2, pp. 181–192, Apr. 2021, doi: 10.22044/jadm.2020.9949.2143.

[21] B. Sabeti, H. Abedi Firouzjaee, A. Janalizadeh Choobbasti, S. H. E. Mortazavi Najafabadi, and A. Vaheb, "MirasText: An Automatically Generated Text Corpus for Persian," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds., Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. Accessed: Dec. 07, 2024. [Online]. Available: https://aclanthology.org/L18-1188.

[22] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard Persian text collection," *Knowledge-Based Systems*, vol. 22, no. 5, pp. 382–387, Jul. 2009, doi: 10.1016/j.knosys.2009.05.002.

[23] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," Mar. 11, 2021, *arXiv*: arXiv:2010.11934. doi: 10.48550/arXiv.2010.11934.

[24] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 01, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

PWNC: پیکره بزرگ‌مقیاس فارسی برای ابهام زدایی و شناسایی موجودیت‌های نام‌دار با استفاده از یادگیری نیمه‌نظارتی و نظارتی

آرش کشت‌کار، سعیده سادات سدیدپور* و حسین شیرازی

دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر، تهران، ایران.

ارسال ۲۰۲۵/۰۴/۰۵؛ بازنگری ۲۰۲۵/۰۶/۰۱؛ پذیرش ۲۰۲۵/۱۱/۰۲

چکیده:

ابهام‌زدایی معنای کلمه از دیرباز یکی از چالش‌های اساسی در پردازش زبان طبیعی بوده است، به‌ویژه در زبان‌هایی مانند فارسی که از غنای ساخت‌واژگانی برخوردارند اما منابع آموزشی محدودی دارند. ابهام ذاتی موجودیت‌های نام‌دار در زبان فارسی، به‌ویژه در متون تخصصی و به دلیل کمبود داده‌های برچسب‌گذاری‌شده، تفسیر معنایی و استخراج اطلاعات را دشوار می‌سازد. پیکره PWNC با هدف رفع چالش‌های همزمان شناسایی موجودیت‌های نام‌دار و ابهام‌زدایی معنای کلمه در زبان فارسی توسعه یافته است. این پیکره اولین پیکره بزرگ‌مقیاس و یکپارچه‌ای که این دو وظیفه را به صورت همزمان پ شتیبانی می‌کند. همچنین با ا ستفاده از یک چارچوب نیمه‌نظارتی و به صورت خودکار ساخته شده و در آن، ترکیبی از معیارهای شباهت‌یابی بافتاری و الگوریتم‌های خوشه‌بندی برای برچسب‌گذاری موجودیت‌های مبهم در ده دسته معنایی به‌کار رفته است. روشی برای دسته‌بندی معنایی واژگان هم‌نویسه در این چارچوب پیشنهاد شده است که عملکردی قوی از خود نشان داده و در مجموع بیش از ۳۰۵ هزار پاراگراف برچسب‌گذاری‌شده، به دقت ٪۸۳، بازیابی ٪۸۱ و نمره F1 معادل ٪۸۲ دست یافته است. تحلیل جامع خطاها نشان داده است که اصلی‌ترین چالش‌ها، تمایز بین معانی بسیار نزدیک و شناسایی موجودیت‌های ضعیف است. با به‌کارگیری راهبردهای تعبیه‌سازی بافتاری، این مشکلات به‌طور قابل توجهی کاهش یافته‌اند.

کلمات کلیدی: ابهام زدایی معنایی کلمه، شناسایی موجودیت‌های نام‌دار، رفع ابهام معنایی، پردازش زبان فارسی.