International Journal of Emerging Research in Computer Science
Volume 2 | Issue 3 | March 2026 | ISSN: 3139-101X (Online)
Available online at: https://ijercs.com/
DOI: https://doi.org/10.59828/ijercs.v2i3.24

Fact Recognition and Extraction Techniques in Natural Language Processing

Prathamesh Pangare✉
Department of Computer Science, Sarhad College of Arts, Commerce & Science, Katraj, Pune, India

Received: 07 March 2026 | **Accepted:** 20 March 2026 | **Published:** 27 March 2026

Abstract

Natural Language Processing has become one of the most useful areas of computer science, especially when it comes to understanding and working with large amounts of text data. One important task within NLP is fact recognition and extraction — which basically means identifying useful facts from unstructured text and organizing them in a way that machines can process. This paper studies the main techniques used for this purpose, including Named Entity Recognition (NER), relation extraction, and Open Information Extraction (OpenIE). The role of modern deep learning models, especially BERT, is also discussed. The study is based on reviewing existing research and published papers in this area. It was found that while deep learning methods have greatly improved the accuracy of fact extraction, there are still some challenges like handling ambiguous language and making these systems work across different domains.

Keywords: *Natural Language Processing, Fact Extraction, Named Entity Recognition, Relation Extraction, Deep Learning.*

1. Introduction

Every single day, millions of documents, news articles, research papers and social media posts are produced in text form. Most of this data is unstructured, meaning it is written in natural human language and not organized in any specific way. For computers and software systems to make use of this data, they need to first understand it and then pull out relevant information from it. This is where Natural Language Processing comes in.

Fact recognition refers to the process of identifying important information — like names, places, dates, events — from a piece of text. Fact extraction then takes it a step further by converting this identified information into structured form, for example as a table or a knowledge graph. Together these two tasks form the backbone of many real-world applications like search engines, chatbots, automated summarization, and question answering systems.

Earlier approaches to this problem used simple rule-based methods and keyword matching, which were limited and didn't work well outside the specific domain they were built for. Over time, with the advancement of machine learning and deep learning, these systems have improved a lot. Today, models like BERT can understand the meaning and context of words much better than older methods. This paper aims to study these developments and understand how fact recognition and extraction has evolved over the years.

2. Literature Review

A lot of research work has been done in the area of fact recognition and extraction. Early work in this field was heavily influenced by the Message Understanding Conferences (MUC) organized in the late 1980s and 1990s. These conferences defined standard tasks like Named Entity Recognition and template filling, and helped researchers compare their systems using common benchmarks. Systems from that era were mostly rule-based, meaning experts manually wrote patterns and rules to identify entities.

Lafferty et al. (2001) introduced Conditional Random Fields (CRFs), which became a very popular approach for sequence labeling tasks like NER. CRFs are statistical models that learn from labeled training data and are better at capturing context compared to earlier methods. Around the same time, Mintz et al. (2009) proposed the concept of distant supervision for relation extraction, where an existing knowledge base like Freebase is used to automatically generate training data from text. This reduced the need for manual annotation, although it introduced some noise in the training data.

Banko et al. (2007) introduced the idea of Open Information Extraction through their TextRunner system, which could extract facts from web text without needing predefined relation types. This approach was later improved by systems like ReVerb and Stanford OpenIE. On the deep learning side, Huang et al. (2015) showed that Bidirectional LSTM combined with CRF (BiLSTM-CRF) gave better NER results than traditional methods by learning richer representations of text.

The biggest shift came with the Transformer architecture (Vaswani et al., 2017) and then BERT (Devlin et al., 2018). BERT is a pre-trained language model that reads text from both directions at once, giving it a much better understanding of word meaning in context. Fine-tuning BERT on NER and relation extraction tasks gave significant improvements over all previous methods. Since then, variants like BioBERT for medical text and SciBERT for scientific papers have shown that domain-specific pre-training can also help a lot.

3. Research Objectives

The main objective of this study is to understand and review the important techniques used for fact recognition and extraction in the field of Natural Language Processing. The study specifically looks at how these techniques have changed over time, from rule-based approaches to modern deep learning models.

Another objective is to compare the performance of different approaches and understand which methods work better and in what conditions. The study also tries to identify the practical challenges that exist in current systems and discuss what future directions look like. Overall, the aim is to get a clear picture of where this field currently stands and what still needs to be worked on.

4. Methodology

This study follows a secondary research approach, meaning no new experiments or datasets were created. Instead, the research is based on reviewing and analyzing existing published work in the area of NLP and information extraction. Sources include academic journals, conference papers from venues like ACL, EMNLP, and NAACL, as well as standard textbooks in the field. Papers were selected based on how relevant they are to the topics of NER, relation extraction, OpenIE, and deep learning for fact extraction. Both older foundational papers and more recent work (up to 2023) were included to get a complete picture of how the field has evolved. The selected papers were then compared and analyzed based on the techniques they use, the datasets they test on, and the results they report.

No primary data collection or coding experiments were performed as part of this study. The findings are therefore based entirely on what has been reported in the reviewed literature. This kind of review-based methodology is common in research papers that aim to give an overview of a technical area rather than introduce a new system or experiment.

4.1 Named Entity Recognition (NER)

NER is the task of identifying and labeling named entities in text, such as names of people, organizations, locations, dates, and so on. It is usually the first step in any fact extraction pipeline. For example, to extract the fact that "Ratan Tata founded Tata Group", the system first needs to identify "Ratan Tata" as a PERSON and "Tata Group" as an ORGANIZATION. On the CoNLL-2003 benchmark, CRF-based systems scored around F1 = 0.87, BiLSTM-CRF models reached ~0.91, and fine-tuned BERT models now achieve around F1 = 0.93.

4.2 Relation Extraction

Relation extraction identifies what semantic relationship exists between two entities in text. For example, from "Barack Obama was born in Hawaii", the extracted triple would be (Barack Obama, bornIn, Hawaii). This is important for building knowledge graphs and answering factual questions. BERT-based models have shown strong performance here too, reaching around F1 = 0.75 on the TACRED benchmark, compared to ~0.60 for earlier feature-based methods.

4.3 Open Information Extraction (OpenIE)

OpenIE systems extract (subject, relation, object) triples from text without needing a predefined set of relation types. Stanford OpenIE and ReVerb are commonly used systems in this category. They use dependency parsing to figure out what goes with what in a sentence. The advantage is that they can work on any domain, but the downside is that the relation phrases they extract are raw text strings, which are harder to use in structured systems without additional processing.

4.4 BERT and Transformer Models

BERT changed NLP in a big way. It is pre-trained on large amounts of text using a masked language modeling objective, which means it learns to predict missing words based on surrounding context — from both sides of the sentence. This bidirectional understanding gives BERT much richer representations than earlier models. When fine-tuned on specific tasks like NER or relation extraction, it outperforms older approaches by a notable margin. Domain-specific versions like BioBERT (medical) and SciBERT (scientific) have extended this advantage to specialized areas.

5. Challenges

Even though the techniques have improved a lot, there are still quite a few challenges that make fact extraction a hard problem in practice. One of the most common issues is language ambiguity. The same word can mean different things depending on context. For example, "Apple" could be a technology company or a fruit. If the surrounding context is not enough to disambiguate, the system might make a wrong extraction.

Another challenge is handling implicit facts. Some sentences contain factual information that is not expressed directly. For example, "He breathed his last in a Mumbai hospital on Friday" is essentially saying someone died, but there is no word like "died" or "death" in the sentence. Most systems are not good at picking up on these kinds of indirect expressions.

Domain portability is also a big issue. A model trained on news articles will often perform poorly on medical reports or legal documents because the language used in these domains is very different. Training separate models for every domain is expensive and not always practical. This is an area where domain-specific pre-trained models like BioBERT help, but they are not available for every domain.

Data requirement is another concern. Deep learning models need large amounts of labeled training data to perform well. Manually annotating text is time consuming and expensive, and for many specialized domains there simply isn't enough labeled data available. Techniques like distant supervision and few-shot learning try to address this, but they come with their own limitations.

Finally, most high-performing models are computationally heavy. Running a BERT-based model requires significant hardware resources, which makes deployment difficult in environments with limited compute — like mobile devices or low-cost servers. This creates a gap between research performance and real-world practical use.

6. Findings

Based on the reviewed literature, it is clear that deep learning based approaches, especially transformer models like BERT, have significantly improved the accuracy of both fact recognition and fact extraction compared to older methods. The improvement is visible across different tasks — NER, relation extraction, and OpenIE all show higher F1 scores with neural approaches than with rule-based or traditional statistical methods.

It was also found that pre-trained language models reduce the dependency on large task-specific labeled datasets. Fine-tuning BERT on a relatively small annotated corpus can still give competitive results, which is a practical advantage for researchers and developers who do not have access to massive datasets.

Domain-specific pre-training was another important finding. Models like BioBERT and SciBERT, which are pre-trained on domain-specific text, consistently outperform general-purpose models on tasks involving specialized language. This shows that the domain of pre-training matters, not just the size of the pre-training corpus.

However, the findings also show that benchmark performance does not always reflect real-world performance. Models that score well on standard test sets sometimes struggle when applied to messier, real-world text. This suggests that evaluation methodology needs to improve, and that robustness testing should be included alongside accuracy metrics in future research.

7. Conclusion

Fact recognition and extraction is a foundational problem in Natural Language Processing, and a lot of progress has been made in solving it over the past two decades. Starting from simple rule-based systems to statistical models like CRFs, and now to transformer-based deep learning models like BERT, each generation of techniques has improved on the previous one in terms of accuracy and flexibility.

The most important development in recent years has clearly been the introduction of pre-trained language models. BERT and its domain-specific variants have shown that learning rich contextual representations from large text corpora and then fine-tuning on specific tasks is a very effective strategy. This has made high-quality fact extraction accessible even without massive labeled datasets.

That said, challenges like language ambiguity, implicit fact extraction, domain portability and computational cost still need to be addressed. Future research directions like few-shot learning, cross-lingual models, and neuro-symbolic approaches look promising. As NLP continues to advance, fact extraction systems will only become more accurate and useful in real-world applications.

References

[1]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2]. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS).

[3]. Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of ICML 2001.

[4]. Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. ACL-IJCNLP 2009.

[5]. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence labeling. arXiv preprint arXiv:1508.01991.

[6]. Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. IJCAI 2007.

[7]. Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Named entity recognition. CoNLL-2003.

[8]. Lee, J., Yoon, W., Kim, S., et al. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.

[9]. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Stanford University.

[10]. Collobert, R., Weston, J., Bottou, L., et al. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12, 2493–2537.

Cite this Article:

Pangare, P. (2026). Fact Recognition and Extraction Techniques in Natural Language Processing. International Journal of Emerging Research in Computer Science, 2(3), 18–21.

Journal URL: *https://ijercs.com/*

DOI: https://doi.org/10.59828/ijercs.v2i3.24