POS Tagging Algorithms for SPO Pairs and Knowledge Graph Extraction

Mayank Tamakuwala
Khoury College of Computer Sciences
Northeastern University
Boston, MA 02115
tamakuwala.m@northeastern.edu

Nithin Bhat
Khoury College of Computer Sciences
Northeastern University
Boston, MA 02115
bhat.nith@@northeastern.edu

Agasti Mhatre
Khoury College of Computer Sciences
Northeastern University
Boston, MA 02115
mhatre.ag @northeastern.edu

Abstract

This project explores optimal algorithms for extracting Subject-Predicate-Object (SPO) pairs from news articles to facilitate knowledge graph construction. We manually generated ground truth SPO pairs and evaluated two different approaches: the Viterbi algorithm (based on Hidden Markov Models) and a fine-tuned BERT model utilizing BIO tagging. Our focus was on token-level labeling and span extraction to transform unstructured news content into structured relational data. The evaluation was conducted using both syntactic and semantic metrics.

Introduction

As information overload increases, the need for automatically structuring unstructured data becomes critical. Knowledge graphs offer a powerful framework for organizing entities and relationships found in natural language. In this project, we focus on converting raw news text into SPO triplets, which are foundational to building knowledge graphs. The challenge lies in the correct identification and labeling of textual spans representing entities and relations. Our research aims to evaluate two contrasting methods — statistical (Viterbi) and deep learning (BERT) — to understand their effectiveness in this domain.

Method

Dataset and Data Pre-processing

We scraped news articles from globenewswire.com across three sectors: finance, healthcare, and technology. Each article was paired with manually curated SPO triplets that served as the ground truth. We tokenized the article content using spaCy, aligned it with labeled triplets using a BIO tagging format, and stored the results for training.

Architecture

We implemented two models for SPO tagging:

Viterbi Algorithm (HMM)

The Viterbi-based model was implemented using the nltk.tag.hmm.HiddenMarkovModelTrainer class. The dataset was first converted into sequences of (word, tag) tuples. The HMM was trained in a supervised manner using annotated SPO triplet sequences. The model learned transition probabilities between tags and emission probabilities from tags to words. At inference, the Viterbi algorithm was applied to find the most likely tag sequence for a given word sequence.

Tagging output was evaluated both at the token level (per tag) and at the sequence level, including semantic similarity with ground truth triplets using ROUGE and BLEU scores. Visualizations such as confusion matrices and error patterns were used for additional insight.

BERT + BIO Tagging

The BERT-based model uses the bert-base-cased pre-trained transformer as a backbone. A token classification layer (linear + softmax) was added to assign one of the seven BIO tags (O, B-SUB, I-SUB, B-PRED, I-PRED, B-OBJ, I-OBJ) to each token.

We used spaCy for text tokenization and BertTokenizerFast to tokenize inputs into subwords, maintaining alignment between tokens and labels. For each token, the model outputs a probability distribution over tag classes, optimized using CrossEntropyLoss with class weights to counter imbalance. Post-processing converts token-level BIO predictions into labeled spans, which are then grouped into triplets within the same sentence using spaCy's sentence segmentation.

Training

The BERT model was fine-tuned using the following parameters:

- Optimizer: AdamW
- Learning rate: 5e-5
- Batch size: 8
- Epochs: 10
- Dropout: 0.1
- Split: 90% training / 10% validation

Each data point was a tuple of tokens, BIO tags, and gold triplets. We used word_ids() from HuggingFace's tokenizer to map subword tokens back to words. During training, the model learned to associate each token with its appropriate BIO label, minimizing the cross-entropy loss while ignoring padding indices (-100). The evaluation included both token-level metrics (precision, recall, F1) and semantic triplet-level metrics (ROUGE, BLEU).

Viterbi-Based Implementation

We used a Hidden Markov Model (HMM) to model tag transitions and emissions over sequences of words in SPO-labeled text. Each word was tagged as one of the labels in a simplified BIO-style

system (e.g., SUBJ, PRED, OBJ). The HMM was trained using NLTK's supervised trainer and evaluated on an 80-20 train-test split. Predictions were compared against ground truth using token-level and sequence-level metrics. Visualizations included tag distribution plots, confusion matrices, and error heatmaps.

Results

BERT Results

The BERT model achieved strong performance across token-level and semantic evaluation metrics:

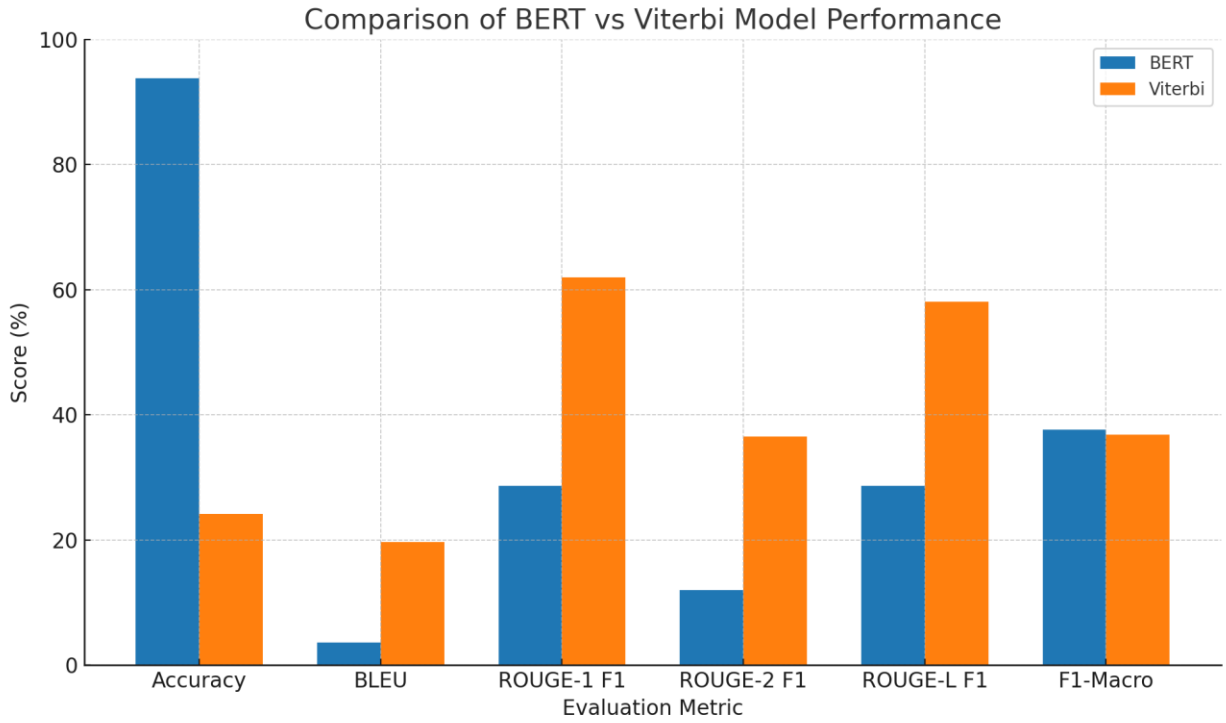
Metric	Score
Accuracy	93.79%
F1-Score (Macro)	37.61%
F1-Score (Weighted)	93.24%
Precision (Macro)	37.80%
Recall (Macro)	41.47%
ROUGE-L (F1)	28.64%
BLEU Score	3.64%

Viterbi Results

Metric	Score
Accuracy	24.11%
BLEU Score	19.71%
ROUGE-1 (F1)	61.97%
ROUGE-2 (F1)	36.57%
ROUGE-L (F1)	58.08%

Tag	Count	Precision	Recall	F1-Score
O	1248	67.44%	2.32%	4.49%
OBJ	219	100.00%	2.74%	5.33%
PRED	381	100.00%	4.20%	8.06%
SUBJ	524	22.58%	99.43%	36.81%

The Viterbi model showed strong recall for the SUBJECT tag (99.43%) but lower precision (22.58%), leading to moderate F1 (36.81%). The BLEU score of 0.1971 was higher than that of the BERT model, suggesting better token-level phrase overlap despite lower overall accuracy. The high ROUGE-1 and ROUGE-L scores confirm partial matches in structure and phrasing, even when exact token alignment was weak.



Discussion

The BERT model's contextual embeddings helped it correctly tag tokens based on sentence meaning and grammatical context. BIO tagging allowed it to segment entity spans accurately even when they appeared across multiple tokens. However, tagging performance dropped slightly when subjects or objects were long or uncommon phrases.

The Viterbi algorithm, while simpler and more interpretable, excelled in recall-heavy tags like SUBJECT, possibly due to transition dominance learned from the dataset. However, it suffered from low precision, especially for OBJECT and PREDICATE classes. This led to inflated false positives and reduced sequence accuracy. Despite these issues, it outperformed BERT in BLEU and ROUGE-1/ROUGE-L scores, highlighting its strength in structure-level sequence generation rather than token classification. In contrast, BERT performed more reliably in fine-grained token labeling but lagged in semantic structure generation when compared to Viterbi.

Conclusion

We demonstrated the potential of deep learning in extracting structured data from unstructured text using BERT with BIO tagging. Our approach provides a foundation for automatic knowledge graph creation. Future work will include implementing and evaluating the Viterbi-based model to offer a complete performance comparison between traditional and deep learning approaches.

Acknowledgment

We would like to thank the open-source community behind *spaCy*, *HuggingFace Transformers*, *sklearn*, and *seqeval*, whose tools and models made this project possible.

Special thanks to Prof. Uzair Ahmad for their guidance.

Code Repository & Data

GitHub: https://github.com/MayankTamakuwala/Knowledge-Graphs/

Data: https://globenewswire.com

References

Introspective Market Research. (2025, February 6). *Knowledge Graph Market Expected To Reach USD 9.23 Billion by 2032, Growing at CAGR 12.45% | Franz Inc., IBM Corporation, Ontotext*. GlobeNewswire News Room; Introspective Market Research. https://www.globenewswire.com/news-release/2025/02/06/3021842/0/en/Knowledge-Graph-Market-Expected-To-Reach-USD-9-23-Billion-by-2032-Growing-at-CAGR-12-45-Franz-Inc-IBM-Corporation-Ontotext.html

Mouna Labiadh. (2024, February 13). *Exploring BERT: Feature extraction & Fine-tuning - DataNess.AI - Medium*. Medium; DataNess.AI. https://medium.com/dataness-ai/exploring-bert-feature-extraction-fine-tuning-6d6ad7b829e7

Naseem, U., Dunn, A. G., Khushi, M., & Kim, J. (2022). Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. *BMC Bioinformatics*, *23*(1). https://doi.org/10.1186/s12859-022-04688-w

OriginTrail. (2024, August 8). *Trend is Your Friend: Knowledge Graphs at the Heart of Gartner's Impact Radar — Here is How the Decentralized Knowledge Graph (DKG) Enhances Reliable AI*. Medium; OriginTrail. https://medium.com/origintrail/trend-is-your-friend-knowledge-graphs-at-the-heart-of-gartners-impact-radar-here-is-how-the-585ce06f087c

Reddit - The heart of the internet. (2024). Reddit.com. https://www.reddit.com/r/LanguageTechnology/comments/1g8brrn/is_pos_tagging_like_with_viterbi_hmm_still_useful/?rdt=57724

SmythOS - Future Trends in AI and Knowledge Graphs. (2024, November 25). SmythOS. https://smythos.com/ai-agents/agent-architectures/knowledge-graphs-and-ai-future-trends/

Team, O. (2024, August 27). *Knowledge graphs on the rise: Gartner's 2024 AI Hype Cycle shows their growing impact*. Ontoforce.com; ONTOFORCE. https://www.ontoforce.com/blog/knowledge-graphs-on-the-rise-gartners-2024-ai-hype-cycle-shows-their-growing-impact

Wiem Souai. (2024, April 5). *Mastering Named Entity Recognition with BERT - UBIAI NLP - Medium*. Medium; UBIAI NLP. https://medium.com/ubiai-nlp/mastering-named-entity-recognition-with-bert-ca8d04b67b18