Text Mining & Chatbots

TP2 Named Entity Recognition and Sentence-level Classification

Ilyes Sais (ilyes.sais@universite-paris-saclay.fr)
Wenchong Pan (wenchong.pan@universite-paris-saclay.fr)

Instructor: Sahar Ghannay

1 Introduction

This second practical session (TP2) focuses on the evaluation of word embeddings and neural architectures on a downstream Natural Language Processing task: Named Entity Recognition (NER). More precisely, we reformulate the NER problem as a sentence-level binary classification task, where each sentence is labeled according to whether it contains at least one named entity or not.

This work builds directly upon the embeddings trained during TP1. We evaluate the impact of different representation and modeling strategies on two corpora of very different nature: a small domain-specific medical corpus (QUAERO FrenchMed) and a large general-domain press corpus (QUAERO FrenchPress). The main objective of this TP is to understand how embedding quality, domain specificity, and model architecture influence performance on a realistic NLP task.

Three modeling approaches are considered: a Long Short-Term Memory network (LSTM) using Word2Vec embeddings, a Convolutional Neural Network (CNN) using FastText embeddings, and a Transformer-based model (CamemBERT) fine-tuned end-to-end.

2 Data Preparation

2.1 Original NER Corpora

The original datasets are provided in a CoNLL-like format, where each line corresponds to a single token annotated with a BIO-style named entity tag. Sentences are separated by empty lines. This format is commonly used for Named Entity Recognition tasks, as it allows fine-grained annotation at the token level.

In the medical corpus (QUAERO FrenchMed), entity categories are highly specialized and include chemical substances (`B-CHEM`), medical procedures (`B-PROC`), anatomical terms (`B-ANAT`), and related biomedical concepts. In contrast, the press corpus (QUAERO FrenchPress) contains more general entity types such as persons, locations, organizations, and products.

Table 1 illustrates an excerpt of the original medical corpus in its raw NER format.

This representation allows precise annotation of entity spans but cannot be used directly for sentence-level classification models without additional preprocessing.

2.2 Sentence Reconstruction and Alignment

The first preprocessing step consists in reconstructing complete sentences by concatenating successive tokens until an empty line is encountered. For each reconstructed sentence, the corresponding list of NER labels is also aggregated.

Table 1: Example of the original medical corpus in CoNLL-like NER format

Token	**NER Label**
PRIALT	B-CHEM
est	O
une	O
solution	B-CHEM
pour	O
perfusion	B-PROC
contenant	O
le	O
principe	B-CHEM
actif	I-CHEM
ziconotide	B-CHEM
.	O

A strict alignment verification is performed to ensure that the number of tokens exactly matches the number of labels for every sentence. This step is essential, as any mismatch would lead to corrupted training instances and unreliable supervision.

For example, the sentence *"PRIALT"* contains a single token and a single label `B-CHEM`. More complex sentences, such as *"Prialt est une solution pour perfusion contenant le principe actif ziconotide"*, contain 23 tokens and 23 corresponding labels. This systematic alignment check guarantees the integrity of the reconstructed data.

2.3 Binary Sentence-level Labeling

To enable a fair comparison between different neural architectures and embedding strategies, the token-level NER task is reformulated as a binary sentence-level classification problem.

Each sentence is assigned a binary label according to the following rule: a sentence receives label 1 if it contains at least one named entity (i.e., at least one label different from `O`); otherwise, it receives label 0. This transformation allows us to focus on the detection of entity-bearing sentences while still reflecting the underlying entity distribution of the corpus.

Table 2 presents an example of the resulting sentence-level dataset after this transformation.

Table 2: Sentence level dataset after binary labeling

ID	**Review (Sentence)**	**Label**
0	PRIALT	1
1	EMEA / H / C / 551	0
2	Qu ' est ce que Prialt ?	1
3	Prialt est une solution pour perfusion contenant le principe actif ziconotide.	1
4	Dans quel cas Prialt est-il utilisé ?	1

This representation is compatible with sentence-level neural classifiers such as LSTM, CNN, and Transformer-based models.

2.4 Dataset Statistics

After preprocessing, the medical corpus is split into three subsets. The training set contains 706 sentences, the validation set contains 649 sentences, and the test set contains 578 sentences.

Each subset follows the same two-column structure, consisting of a sentence text and a binary label.

The same preprocessing pipeline is applied to the press corpus, which results in significantly larger datasets due to the higher volume and diversity of journalistic text. Using an identical pipeline for both corpora ensures that performance differences observed during evaluation are attributable to modeling choices and domain characteristics rather than preprocessing discrepancies.

3 Embedding-based Neural Models

3.1 Word2Vec + LSTM Model

The first neural architecture is based on a Long Short-Term Memory network. Sentences are tokenized and mapped to integer sequences using a vocabulary built exclusively on the training data. This vocabulary is explicitly saved to disk to ensure reproducibility and correct evaluation at test time.

Each token is mapped to a 100-dimensional Word2Vec embedding trained during TP1. The LSTM processes the sequence and the last hidden state is used as a sentence representation. A fully connected layer followed by a sigmoid activation outputs a probability indicating the presence of named entities.

3.2 FastText + CNN Model

The second architecture relies on a Convolutional Neural Network combined with FastText embeddings. Unlike Word2Vec, FastText incorporates subword information, making it more robust to morphological variations and rare words.

The CNN applies multiple convolution filters with different kernel sizes to capture local n-gram patterns. The resulting features are pooled and passed to a fully connected layer to perform binary classification.

As with the LSTM model, the vocabulary used for encoding is saved during training to allow consistent evaluation.

4 Transformer-based Model: CamemBERT

The third approach uses CamemBERT, a Transformer-based language model pre-trained on large French corpora. Unlike the previous models, CamemBERT does not rely on static word embeddings but learns contextual representations dynamically.

The model is fine-tuned directly on the sentence classification task using the HuggingFace `Trainer` API. Tokenization, attention masks, and padding are handled automatically. Evaluation metrics are computed at the end of each epoch, and the best model is selected based on validation performance.

5 Semantic Similarity Evaluation

The quality of the learned word embeddings is evaluated through a semantic similarity analysis based on cosine similarity. For a given target word, the objective is to retrieve the most semantically related words according to the embedding space. This qualitative evaluation provides valuable insight into how well each model captures semantic, morphological, and domain-specific information.

5.1 Evaluation Words

Five representative words were selected to cover both domain-specific and general semantic phenomena. The words *patient*, *treatment*, *disease*, and *solution* are central to the medical domain and are expected to reveal the ability of embeddings to capture specialized biomedical semantics. The word *yellow* was intentionally included as a general and ambiguous term in order to analyze how models behave outside of a specialized domain and how they rely on contextual versus morphological cues.

5.2 Evaluation Methodology

Two complementary evaluation methods were employed. First, the `most_similar` function from the `gensim` library was used to directly retrieve the top-10 closest words according to cosine similarity in the embedding space. Second, cosine similarity was computed explicitly on the vector representations to confirm the consistency of the results. Both methods yielded equivalent rankings, confirming the robustness of the evaluation procedure.

5.3 Medical Corpus Analysis

On the medical corpus, clear differences emerge between embedding approaches. Word2Vec CBOW embeddings tend to emphasize frequent contextual words. For instance, the nearest neighbors of *patient* include highly frequent tokens such as *cette*, *le*, and *qui*. Similarly, *traitement* is associated with both medical terms such as *médecin* and *TYSABRI*, but also with function words such as *que*. This behavior reflects the sensitivity of the CBOW model to frequent contextual patterns, particularly in relatively small corpora.

In contrast, the Word2Vec Skip-gram model trained on the medical corpus produces more semantically focused representations. The word *traitement* is strongly associated with medically relevant concepts such as *cancer*, *Parkinson*, and *expérimenté*. Likewise, *maladie* is linked to terms such as *Parkinson*, *liée*, and *charge*, which reflect meaningful medical relations rather than purely syntactic proximity. This confirms the well-known advantage of Skip-gram for capturing informative and less frequent domain-specific vocabulary.

FastText CBOW embeddings exhibit particularly strong behavior on the medical corpus. The nearest neighbors of *traitement* include morphological variants such as *Traitement*, *traitements*, and even misspelled forms like *Taaitement*, all with near-perfect similarity scores. Similarly, *patient* is associated with *Patient*, *patiente*, and medically relevant terms such as *tremblements* and *pansements*. The word *solution* is linked to *dilution*, *dissolution*, and *flacon*, which are directly relevant to pharmaceutical preparation contexts. These results highlight the benefit of FastText's subword modeling, which captures morphological structure and capitalization variations while remaining robust to rare or noisy forms.

5.4 Press Corpus Analysis

The behavior of embeddings trained on the press corpus differs markedly. Word2Vec CBOW embeddings reflect more abstract and general semantics. For example, *traitement* is associated with words such as *coût*, *financement*, and *système*, reflecting socio-economic discourse rather than medical meaning. The word *jaune* is primarily linked to sports-related or named entities such as *maillot* or *Perrot*, which is consistent with journalistic usage.

Word2Vec Skip-gram embeddings on the press corpus provide slightly more focused semantics. The word *maladie* is associated with specific diseases such as *Alzheimer* and *grippe*, while *solution* is linked to abstract concepts such as *alternative* and *consensuelle*. Nevertheless, the semantic specificity remains lower than in the medical corpus due to the broader thematic scope of press data.

FastText CBOW embeddings trained on the press corpus show a clear tendency toward morphological similarity rather than semantic specialization. For instance, *patient* is associated with *patientent*, *impatient*, and verbs such as *ratifient* or *trient*, which share character-level patterns but are not semantically equivalent. Similarly, *solution* retrieves *révolution*, *résolution*, and *évolution*, demonstrating that FastText prioritizes subword overlap in large, general-domain corpora.

5.5 Comparative Discussion

Across all experiments, the impact of the training corpus is more pronounced than the choice of embedding algorithm. Medical embeddings consistently produce more relevant and domain-specific neighbors for biomedical terms than press embeddings. Skip-gram outperforms CBOW in capturing specialized semantics, particularly for less frequent and informative terms. FastText proves extremely effective in the medical domain due to its ability to leverage morphological information, while in the press domain it tends to emphasize surface-level similarity over deep semantics.

5.6 Implications for Named Entity Recognition

These observations strongly suggest that embeddings trained on domain-matched data are crucial for downstream tasks such as Named Entity Recognition. In particular, medical FastText and Skip-gram embeddings are expected to provide substantial gains in detecting rare entities, handling spelling variations, and maintaining semantic coherence. These hypotheses are empirically validated in the second practical session through quantitative NER experiments.

==========

6 Results

6.1 Medical Domain Results

Model	Precision	Recall	F1-score	Accuracy	Domain
LSTM (Word2Vec)	0.8475	1.0000	0.9174	0.8475	Medical
CNN (FastText)	0.8587	0.9945	0.9217	0.8567	Medical
CamemBERT	0.8910	0.9800	0.9424	0.8910	Medical

The results clearly show that CamemBERT achieves the best overall performance on the medical corpus. Its contextual representations allow it to capture complex medical terminology more effectively than static embeddings. The CNN with FastText embeddings performs slightly better than the LSTM, confirming the importance of subword modeling in medical texts.

6.2 Press Domain Results

Model	Precision	Recall	F1-score	Accuracy	Domain
LSTM (Word2Vec)	0.8755	0.8873	0.8814	0.8251	Press
CNN (FastText)	0.8637	0.8888	0.8761	0.8159	Press
CamemBERT	0.8871	0.9679	0.9258	0.9160	Press

On the press corpus, CamemBERT again outperforms embedding-based models by a significant margin. The difference in accuracy is particularly noticeable, highlighting the benefit of contextualized representations in heterogeneous and large-scale text data.

7 Discussion

Across both domains, Transformer-based models consistently outperform embedding-based neural networks. This can be explained by CamemBERT's ability to capture long-range dependencies and contextual meaning, which is crucial for identifying named entities.

Among embedding-based approaches, FastText combined with CNN generally performs better than Word2Vec combined with LSTM, especially in the medical domain. This confirms observations from TP1 regarding the importance of subword information and morphological robustness.

The impact of domain is also clearly visible. All models achieve higher recall in the medical corpus, where entity patterns are more systematic, whereas performance on the press corpus is more challenging due to greater lexical diversity.

8 Conclusion

In this TP, we demonstrated how different embedding strategies and neural architectures affect performance on a Named Entity Recognition-related task. The experiments confirm that domain-specific embeddings are beneficial, but that contextualized Transformer models provide the most robust and accurate solution.

CamemBERT emerges as the best-performing model in both domains, making it the most suitable choice for real-world NER applications. However, embedding-based models remain valuable when computational resources are limited or when interpretability is required.

This work illustrates the full pipeline from raw NER annotations to sentence-level classification and provides a clear comparison of modern NLP approaches in a controlled experimental setting.