Text Mining & Chatbots

TP1 Report: Word Embeddings Training and Analysis

Professor: Sahar Ghannay

Group Members:
Ilyes Sais
ilyes.sais@universite-paris-saclay.fr
Wenchong Pan
wenchong.pan@universite-paris-saclay.fr

1 Introduction

Word embeddings play a central role in modern Natural Language Processing (NLP) systems by providing dense vector representations of words that capture semantic and syntactic information. The objective of this practical session (TP1) is to build and compare several word embedding models trained on corpora from different domains and sizes, and to analyze their semantic behavior.

Specifically, we trained and analyzed word embeddings using:

- Word2Vec Continuous Bag-of-Words (CBOW)
- Word2Vec Skip-gram
- FastText CBOW

These embeddings were trained on:

- A **medical domain corpus** (QUAERO_FrenchMed), small-sized but domain-specific
- A **general news corpus** (QUAERO_FrenchPress), large-sized and non-medical

The learned embeddings will later be reused in TP2 for Named Entity Recognition (NER).
—

2 Resources and Tools

2.1 Corpora

Two corpora were used:

- **QUAERO_FrenchMed**: a specialized medical corpus
- **QUAERO_FrenchPress**: a large non-medical press corpus

Each corpus is provided in a “one sentence per line” format, with tokens separated by spaces.

2.2 Libraries

The following Python libraries were used:

- `gensim` for Word2Vec models
- `fasttext` for FastText embeddings

All experiments were conducted using Python, mainly on Google Colab.

—

3 Word Embeddings Training

In this section, we describe the training process of the word embedding models and provide a detailed qualitative analysis of their semantic behavior. A total of **six word embedding models** were trained using two different corpora and three embedding approaches.

3.1 Overview of Trained Models

The following embedding models were trained:

- Word2Vec CBOW trained on the medical corpus (QUAERO_FrenchMed)
- Word2Vec Skip-gram trained on the medical corpus
- FastText CBOW trained on the medical corpus
- Word2Vec CBOW trained on the press corpus (QUAERO_FrenchPress)
- Word2Vec Skip-gram trained on the press corpus
- FastText CBOW trained on the press corpus

All models were trained using the same hyperparameters:

- Embedding dimension: 100
- Minimum word frequency: 1

This controlled setup allows a fair comparison between embedding approaches and between training corpora.

—

3.2 Word2Vec Models on the Medical Corpus

3.2.1 CBOW on the Medical Corpus

The Word2Vec CBOW model trained on the medical corpus exhibits a strong domain-specific behavior, despite the relatively small size of the corpus.

Examples of nearest neighbors include:

- **“traitement”** → médecin, TYSABRI, devra
- **“solution”** → poudre, flacon, diluer
- **“maladie”** → administration, évolution

These neighboring words are clearly related to:

- medical procedures,
- pharmaceutical products,
- treatment instructions.

However, CBOW also retrieves highly frequent function words (such as *que*, *la*, or *du*), which introduces some semantic noise. This behavior is expected, as CBOW relies heavily on local context and is sensitive to frequent words, especially in small corpora.

—

3.2.2 Skip-gram on the Medical Corpus

The Skip-gram model provides more focused and informative semantic representations than CBOW.

Notable examples include:

- **“traitement”** → cancer, Parkinson, expérimenté
- **“solution”** → injectable, perfusion, intraveineuse
- **“maladie”** → Parkinson, charge, liée

Compared to CBOW:

- Skip-gram captures rarer but more meaningful medical terms
- It models specialized medical vocabulary more accurately

This result is consistent with the theoretical properties of Skip-gram, which is known to perform better on infrequent and domain-specific words.

—

3.3 FastText CBOW on the Medical Corpus

FastText shows particularly strong performance on the medical corpus.

Key observations include:

- Very high cosine similarity scores (often close to 1.0)
- Strong sensitivity to morphological variations

Examples:

- **“traitement”** → Traitement, traitements, Allaitement

- **“patient”** → patiente, Patient
- **“solution”** → dilution, dissolution

These results highlight the advantages of FastText:

- use of character n-grams,
- robustness to capitalization,
- ability to handle rare or unseen words.

As a result, FastText is particularly well-suited for medical texts, where terminology is complex and morphologically rich.

—

3.4 FastText CBOW on the Press Corpus

On the large press corpus, FastText mainly captures morphological similarity rather than domain-specific semantics.

Typical examples include:

- **“patient”** → impatient, patientent
- **“solution”** → révolution, résolution
- **“jaune”** → lune, brune, jeune

Although these neighbors are linguistically valid, they are:

- less semantically meaningful for Named Entity Recognition,
- often based on suffix or character overlap.

This confirms that FastText emphasizes form-level similarity, especially in large and general-domain corpora.

—

3.5 Global Comparison of Embedding Approaches

Model	Medical Corpus	Press Corpus
Word2Vec CBOW	Contextual but noisy	General and abstract
Word2Vec Skip-gram	Strong medical semantics	Better than CBOW
FastText CBOW	Excellent semantics + morphology	Mainly morphological

3.5.1 Key Findings

- The training corpus has a stronger impact than the embedding model itself
- Skip-gram outperforms CBOW for specialized vocabulary
- FastText is the most robust approach for rare words and morphology

—

3.6 Implications for Named Entity Recognition (TP2)

Based on these qualitative analyses, we expect the following outcomes for TP2:

- Medical embeddings to outperform press embeddings on medical NER
- FastText medical embeddings to be particularly effective for:
  - rare entities,
  - complex terminology,
  - spelling variations
- Skip-gram medical embeddings to provide strong semantic consistency

These hypotheses will be quantitatively validated during the Named Entity Recognition experiments in TP2.

3.7 Training Configuration

The following hyperparameters were used for all models:

- Embedding dimension: 100
- Minimum word frequency (`min_count`): 1
- Context window: default values

Each model and its corresponding word vectors were saved for later analysis and reuse.

—

4 Semantic Similarity Evaluation

In order to assess the quality and usefulness of the learned word embeddings, we conducted a semantic similarity analysis based on cosine similarity. The objective of this evaluation is to examine whether words that are semantically or contextually related are located close to each other in the embedding space.

For each embedding model, a target word was selected and its vector representation was compared to all other word vectors in the vocabulary. The similarity between two words was measured using cosine similarity, which evaluates the angle between their vector representations. For each target word, the ten most similar words were retrieved, allowing a qualitative inspection of the semantic coherence captured by each embedding approach.

This evaluation was carried out independently for all embedding models trained during this practical session, namely Word2Vec CBOW, Word2Vec Skip-gram, and FastText CBOW, on both the medical and press corpora. By comparing the nearest neighbors returned by each model, we are able to analyze the impact of the embedding approach as well as the influence of the training corpus on the learned representations.

The evaluation focused on a fixed set of representative words: *patient*, *treatment*, *disease*, *solution*, and *yellow*. These words were deliberately chosen to cover different semantic categories.

Some of them are strongly domain-specific and highly relevant to the medical field, while others are more general-purpose words. This choice enables a clear comparison between embeddings trained on medical data and those trained on non-medical press data.

Two complementary evaluation methods were used to compute semantic similarity. First, we relied on the `most_similar` function provided by the `gensim` library, which directly returns the nearest neighbors of a given word based on cosine similarity in the embedding space. This method is convenient and efficient, as it operates directly on the trained models and their internal vector representations.

In parallel, we also computed cosine similarity explicitly using vector representations. In this case, the embedding vectors were loaded from the saved models, and cosine similarity was calculated manually to retrieve the most similar words. This second approach provides greater transparency and allows direct control over the similarity computation, while serving as a validation of the results obtained with the `gensim` method.

Using these two complementary approaches ensures the robustness of the evaluation and confirms that the observed similarities are intrinsic to the learned embeddings rather than artifacts of a specific implementation. The results of this semantic similarity analysis are presented and discussed in the following sections, where we compare embedding approaches on the same corpus and analyze the impact of training data across different domains.

4.1 Evaluation Words

The following words were selected for evaluation:

- patient
- treatment
- disease
- solution
- yellow

4.2 Evaluation Methods

Two approaches were used:

- **Gensim `most_similar`** method
- Cosine similarity using vector representations

—

5 Results Analysis

5.1 Impact of Embedding Method (Same Corpus)

On the same corpus, the choice of embedding method significantly impacts semantic quality:

- Skip-gram generally produces richer semantic neighborhoods than CBOW
- FastText captures subword information, leading to better handling of rare and morphologically complex words

This behavior is especially visible in the medical corpus, where specialized terminology benefits from subword modeling.

5.2 Impact of Corpus Type (Same Method)

When comparing the same embedding method across corpora:

- Medical embeddings retrieve domain-specific terms (e.g., clinical vocabulary)
- Press embeddings retrieve more general and contextual words

This confirms the strong influence of domain and corpus size on embedding quality.

—

6 Discussion

The experiments demonstrate that:

- Domain-specific corpora are essential for specialized tasks
- Larger corpora improve general semantic coverage
- FastText provides robustness for rare and unseen words

These observations justify the use of multiple embedding strategies depending on downstream tasks such as NER.

—

7 Conclusion

In this TP, we successfully trained and analyzed six word embedding models using Word2Vec and FastText on medical and non-medical corpora. The semantic similarity analysis highlighted the impact of both the embedding method and the training corpus.

These embeddings constitute a crucial foundation for the Named Entity Recognition experiments conducted in TP2.

—