Normalization of Hebrew Medical Text Using Large Language Models

Eden Faingold, Linor Cohen, Karin Vashdi

29 October 2024

Abstract

Normalization of unstructured medical text, especially in Hebrew, remains a challenging task for machine learning models. Here, we used an approach of joint named entity recognition (NER) and demonstrate how pretrained large language models (DictaLM 2.0) can be fine-tuned to normalize complex Hebrew medical transcripts. We focus on head MRI reports, testing two main tasks: structuring the reports into standardized sections, and organizing sentences into paragraphs related to specific anatomical regions. Using structured transcripts for sentence extraction and labeling, we employ section titles mapping and hierarchical anatomical regions mapping for fine-tuning the model. The fine-tuned model is capable of transforming unstructured Hebrew radiological reports into structured formats, effectively organizing the content according to medical standards. This approach represents a simple, accessible, and highly flexible method for normalizing unstructured medical texts, enabling efficient conversion of Hebrew transcripts into structured documents.

Introduction

In the radiology department at Ichilov Hospital, data standardization and normalization of Speech-To-Text generated medical transcripts pose significant challenges. When doctors forget to mention details about specific body parts while recording, they must stop, manually locate the relevant section in the system, add the details, and then resume recording. This process, along with the manual editing needed to normalize the transcript's structure, is time-consuming and inefficient. To address these issues, we present an NLP-driven approach using an advanced Hebrew large language model, DictaLM2.0, to automate the structuring of radiological transcripts. By transforming semi-structured head MRI scan reports into fully structured files, our goal is to facilitate information extraction and improve data accessibility for downstream medical applications.

Recent advances in natural language processing (NLP), particularly in the development of large language models, have shown promise for structuring unstructured medical information. In the medical domain, most efforts have focused on named entity recognition (NER) to label important medical entities, such as anatomical regions or findings, within clinical texts. However, these approaches often require multiple steps and manual post-processing, limiting their efficiency and scalability for complex medical reports like radiological transcripts. With the introduction of more sophisticated language models, such as DictaLM2.0, we see an opportunity to leverage these advancements to automate the normalization and structuring of medical transcripts generated by Speech-To-Text systems, reducing manual effort required and increasing the utility of radiological reports.

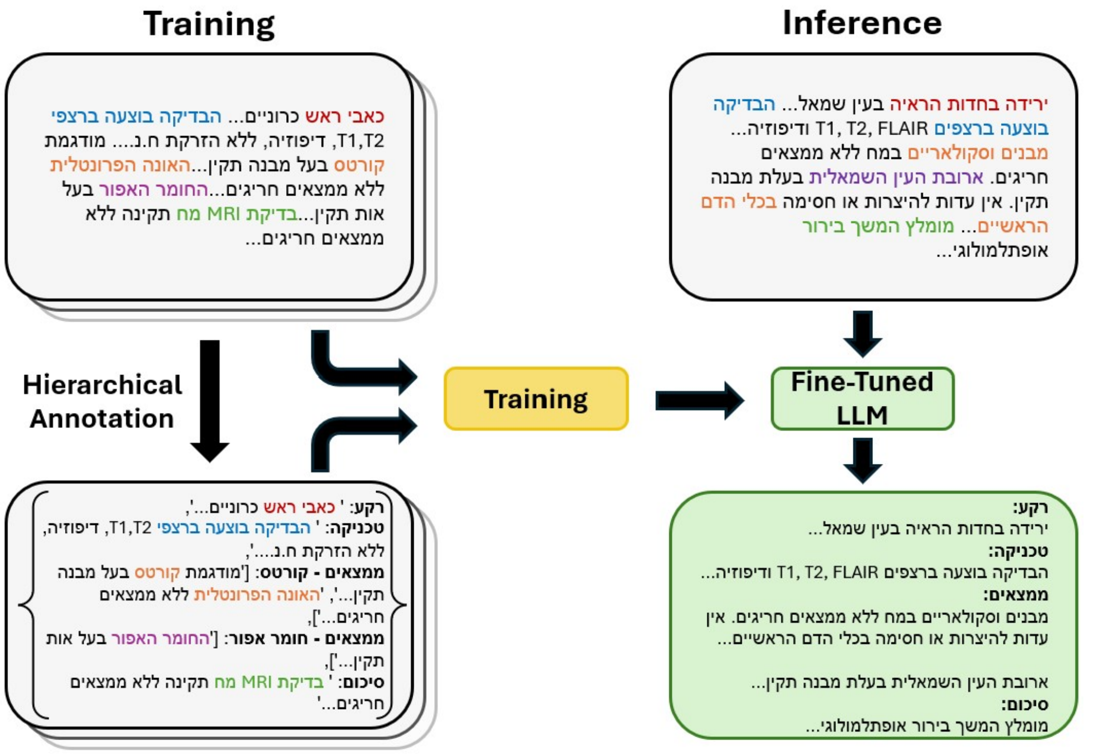
Even though other large language models (LLMs) have demonstrated similar or superior capabilities in English, DictaLM 2.0 stands out as one of the best for Hebrew. According to benchmark comparisons, DictaLM 2.0 consistently achieved top scores across various NLP tasks in Hebrew, including question answering, sentiment analysis, and translation, outperforming other open models like Google Gemma-7B and Llama-2-7B. This strong performance made DictaLM 2.0 a suitable choice for our project. By fine-tuning DictaLM 2.0, we leveraged its robust language capabilities to automate the normalization and structuring of head MRI reports, providing a practical solution specifically tailored to Hebrew medical texts.

Table 1: Hebrew Benchmark for LLMs, Published by Dicta:

Model	**Average**	**QA TLNLS (HeQ)**	**Sentiment Acc**	**Winograd (Binary)**	**Translation**
`dictalm2.0-instruct`	59.48	76.9	56.3	69.42	35.3
`dictalm2.0`	59.46	75.69	59.67	66.91	35.57
`gemma-7b`	57.28	75.86	62.73	62.59	27.93
`Gemma-11B-V2`	54.21	66.3	62.5	60.79	27.25
`Mistral-7B`	53.95	73.52	48.57	64.75	28.98
`Meta-llama-8B`	53.02	72.66	51.8	60.79	26.85
`gemma-2b`	49.31	68.39	61.73	47.48	19.65
`Mistral-7B-v0.1`	41.92	65.81	34.97	51.44	15.48
`Llama-2-7b-hf`	38.58	52.51	40.13	51.44	10.22

To address the challenges of structuring radiology reports, we employed a joint named entity recognition (NER) approach using DictaLM 2.0. Our goal was to normalize head MRI transcripts by structuring them into standardized sections (e.g., "Background," "Findings") and further organizing the "Findings" section into paragraphs based on specific anatomical regions (e.g., "Meninges," "Brain stem"). To accomplish this, we developed section title mapping and hierarchical anatomical regions mapping, which allowed us to create domain-specific labels for the transcripts. By fine-tuning DictaLM 2.0 on this labeled dataset, we automated the structuring process, enabling the model to transform semi-structured text into fully structured reports without requiring extensive manual intervention.

Fig. 1: Overview of the proposed approach to document-level named entity recognition for text normalization task



כאבי ראש כרוניים... הבדיקה בוצעה ברצפי
T1,T2, דיפוזיה, ללא הזרקת ח.נ.... מודגמת
קורטס בעל מבנה תקין...האונה הפרונטלית
ללא ממצאים חריגים...החומר האפור בעל
אות תקין...בדיקת MRI מח תקינה ללא
ממצאים חריגים...

ירידה בחדות הראיה בעין שמאל... הבדיקה
בוצעה ברצפים T1, T2, FLAIR ודיפוזיה...
מבנים וסקולאריים במח ללא ממצאים
חריגים. ארובת העין השמאלית בעלת מבנה
תקין. אין עדות להיצרות או חסימה בכלי הדם
הראשיים... מומלץ המשך בירור
אופתלמולוגי'...

רקע: ' כאבי ראש כרוניים...',
טכניקה: ' הבדיקה בוצעה ברצפי T1,T2, דיפוזיה,
ללא הזרקת ח.נ....',
ממצאים - קורטס: ['מודגמת קורטס בעל מבנה
תקין...', 'האונה הפרונטלית ללא ממצאים
חריגים...'],
ממצאים - חומר אפור: ['החומר האפור בעל אות
תקין...'],
סיכום: ' בדיקת MRI מח תקינה ללא ממצאים
חריגים...'

רקע:
ירידה בחדות הראיה בעין שמאל...
טכניקה:
הבדיקה בוצעה ברצפים T1, T2, FLAIR ודיפוזיה...
ממצאים:
מבנים וסקולאריים במח ללא ממצאים חריגים. אין
עדות להיצרות או חסימה בכלי הדם הראשיים...

ארובת העין השמאלית בעלת מבנה תקין...
סיכום:
מומלץ המשך בירור אופתלמולוגי'...

The primary objective of this study was to automate the normalization and structuring of semi-structured head MRI radiological transcripts to improve data accessibility and facilitate information extraction. To achieve this, we fine-tuned DictaLM 2.0 using a dataset of manually pre-processed and labeled transcripts. The methodology involved several key steps: first, preprocessing the text by cleaning, tokenizing, and labeling sentences using section titles and hierarchical head regions mappings. We employed a section titles mapping to standardize section headers and a hierarchical head regions mapping to label anatomical entities within the "Findings" section. These labeled datasets were then used to fine-tune DictaLM 2.0 with the aim of transforming the unstructured text into a fully standardized format, with each sentence assigned to its correct section and anatomical region. This approach demonstrates the effectiveness of using Hebrew LLMs to overcome the challenges of structuring radiological texts.

Results

The performance of the fine-tuned DictaLM 2.0 model was evaluated against the original DictaLM 2.0 model using F1 score and Positive Match Score. The F1 score was calculated based on two key tasks: structuring reports into standardized sections and organizing sentences into paragraphs related to anatomical regions. We used a test set of 100 text files for this evaluation - 100 unstructured files alongside the corresponding 100 manually edited structured files for reference, to evaluate the model outputs.

Structuring Reports into Standardized Sections

The task of structuring radiological reports into standardized sections was evaluated using precision, recall, and F1 score metrics. The DictaLM 2.0 model achieved a precision of 0.78, a recall of 0.74, and an F1 score of 0.76. After fine-tuning, the model demonstrated a precision of 0.90, a recall of 0.86, and an F1 score of 0.88, indicating significant improvements in structuring reports into their respective sections.

Table 2: F1 Scores for Section Labeling (NER for document sections)

Model	**Precision**	**Recall**	**F1 Score**
DictaLM 2.0	0.78	0.74	0.76
Fine-Tuned DictaLM 2.0	0.90	0.86	0.88

Organizing Sentences into Anatomical Paragraphs

For the task of organizing sentences within the "Findings" section into paragraphs related to specific anatomical regions, the DictaLM 2.0 model achieved a precision of 0.75, recall of 0.69, and an F1 score of 0.72. The fine-tuned model achieved a precision of 0.87, recall of 0.83, and an F1 score of 0.85, reflecting its enhanced capability to accurately group sentences by anatomical context.

Table 3: F1 Scores for Anatomical Region Labeling (NER for anatomy-based grouping)

Model	**Precision**	**Recall**	**F1 Score**
DictaLM 2.0	0.75	0.69	0.72
Fine-Tuned DictaLM 2.0	0.87	0.83	0.85

Positive Match Score

The Positive Match Score was used as an overall assessment of the document structure, providing a binary metric (0 or 1) to indicate whether the entire document was correctly structured in terms of section assignment and paragraph grouping. Flexibility was allowed in the ordering of sentences within paragraphs. The original DictaLM 2.0 model had a Positive Match Score of 0.51, while the fine-tuned model achieved a Positive Match Score of 0.73.

Table 4: Positive Match Score for Structuring Radiological Reports

Model	**Positive Match Score**
DictaLM 2.0	0.51
Fine-Tuned DictaLM 2.0	0.73

Discussion

Overall, the approach of fine-tuning DictaLM 2.0 proved effective in normalizing and structuring Hebrew radiological transcripts into standardized sections and anatomical paragraphs. The model demonstrated the capability to transform unstructured medical text into a consistent, organized format, specifically addressing the complexity inherent in head MRI reports. We hope this approach enables more efficient conversion of Hebrew transcripts into structured documents, ultimately improving the clinical workflow.

Processing Hebrew medical text posed unique challenges, mainly due to the lack of suitable off-the-shelf tools and libraries. Unlike more widely supported languages like English, there are limited options for Hebrew-specific NLP libraries. This necessitated custom solutions for sentence tokenization and labeling. Despite these challenges, using a specialized Hebrew LLM like DictaLM 2.0 allowed us to overcome many of the language-specific complexities in the transcripts.

One of the major limitations of the current study was the relatively small size of the dataset, which led to overfitting during model training. The fine-tuned model performed well on familiar anatomical regions included in the training data but struggled to generalize to less common or unseen regions. This highlights the need for a larger and more diverse dataset to enhance the model's robustness and generalizability across different types of radiological reports.

Future work will focus on expanding the dataset to cover different types of radiological reports beyond head MRI, which will help improve the model's generalizability and reduce overfitting. More importantly, the LLM can be utilized to automate the annotation process. Initially, the head regions mapping was created manually, which limits scalability when expanding to larger datasets. To address this, we suggest using an approach of joint named entity recognition and relation extraction (NERRE). This approach would streamline the grouping of related sentences, making the annotation process more efficient as we scale to larger and more complex datasets.

Methods

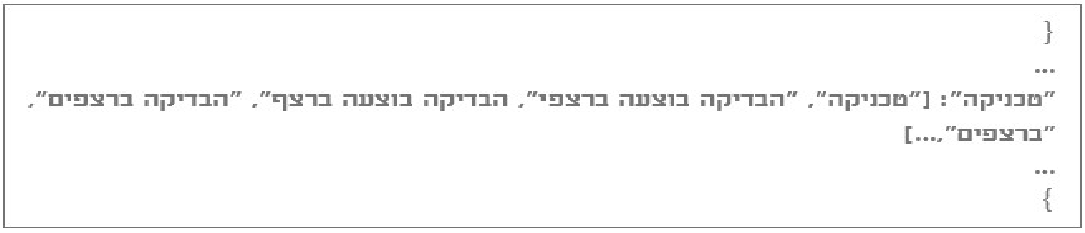
Text Cleaning

To prepare the radiological transcripts for information extraction, we began by cleaning the raw text files. Each transcript contained extraneous elements such as headers, footers, and XML-like tags, which were removed using regular expressions. The goal was to eliminate noise and retain only the clinically relevant content for further processing. Special attention was given to Hebrew-specific encoding issues, ensuring proper handling of unique characters and punctuation.

Data Labeling and Named Entity Recognition (NER)

Radiological transcripts often contained inconsistencies in section titles, with different terminologies used across different files and, in some cases, no explicit section titles at all. To address this, we used Section Titles Mapping, which involved a word normalization technique to map various possible titles to a standardized set of sections: "Background," "Technique," "Findings," and "Summary." This mapping, stored in a JSON file, allowed us to label each sentence in the transcript to a specific section, regardless of variations in the original headings. This ensured that the entire text was parsed and organized consistently for downstream tasks, even in cases where titles were missing or inconsistent.

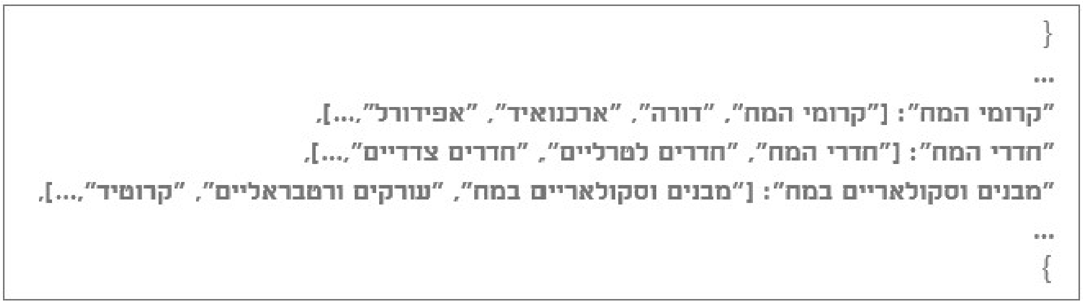
Fig. 2: Excerpt from the Section Titles Mapping JSON:



```
{
...
”טכניקה”: [”טכניקה”, ”הבדיקה בוצעה ברצפי”, הבדיקה בוצעה ברצף”, ”הבדיקה ברצפים”,
”ברצפים”,...]
...
}
```

The "Findings" section in the transcripts required further structuring based on anatomical references. We developed a Hierarchical Head Regions Mapping stored in a JSON file, which included main anatomical regions and their respective sub-regions. This hierarchical mapping was used to label sentences in the "Findings" section by relating mentions of sub-regions to their broader anatomical categories, forming composite labels like "Findings - Meninges." This approach effectively served as a domain-specific Named Entity Recognition (NER) system for head MRI transcripts. Due to the complexity of Hebrew medical terminology and the lack of suitable libraries like SpaCy for Hebrew, we relied on a custom implementation for this task.

Fig. 3: Excerpt from the Hierarchical Head Regions Mapping JSON:



```
{
...
"קרומי המח": ["קרומי המח", "דורה", "ארכנואיד", "אפידורל",...],
"חדרי המח": ["חדרי המח", "חדרים לטרליים", "חדרים צדדיים",...],
"מבנים וסקולאריים במח": ["מבנים וסקולאריים במח", "עורקים ורטבראליים", "קרוטיד",...],
...
}
```

Model Training

The labeled dataset was used to fine-tune the DictaLM 2.0 language model for our specific sequence classification and named entity recognition (NER) task. We accessed DictaLM 2.0 through Hugging Face and used its tokenizer to tokenize the Hebrew sentences, ensuring compatibility with Hebrew text structure and punctuation. We used the Transformers approach for fine-tuning the model, leveraging PyTorch to implement a sequence classification model capable of transforming semi-structured input into structured, labeled output. The model was trained to classify each sentence into its respective section and anatomical category, effectively normalizing the transcripts. The training workflow was implemented using the Transformers library's Trainer class, and the fine-tuned model was subsequently saved for deployment on new radiological transcripts.

Evaluation Criteria

The evaluation of the fine-tuned model was conducted using F1 score and Positive Match Score, calculated as follows:

F1 Score

The F1 score was calculated for each of the two key tasks related to the normalization. Structuring Reports into Standardized Sections:

$$\text{Precision} = \frac{\text{No. of correct section labels retrieved}}{\text{No. of section labels retrieved}}$$

$$\text{Recall} = \frac{\text{No. of correct section labels retrieved}}{\text{No. of section labels in test set}}$$

Organizing Sentences into Paragraphs Related to Anatomical Regions:

$$\text{Precision} = \frac{\text{No. of correct anatomical labels retrieved}}{\text{No. of anatomical labels retrieved}}$$

$$\text{Recall} = \frac{\text{No. of correct anatomical labels retrieved}}{\text{No. of anatomical labels in test set}}$$

The F1 score for both tasks is calculated as the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Averaging Method

Macro averaging was used to calculate precision, recall, and the resulting F1 scores across multiple sections and anatomical categories, ensuring balanced evaluation regardless of category frequency in the dataset.

Positive Match Score

A modified version of the Exact Match criterion was used to assess the overall correctness of the structured output. The Positive Match Score is a binary metric (0 or 1), indicating whether the entire document is an exact match to the reference test document, regardless of the order of sentences within each paragraph in the findings section.

$$\text{Positive Match} = \frac{\text{No. of Documents with Positive Match}}{\text{No. of Documents in Test Set}}$$

Data Availability

All data used in this project are confidential medical records belonging to Ichilov Hospital and cannot be publicly shared.

Code Availability

The code used for this study is available at:
`https://github.com/edenfaingold/Hebrew-Medical-Text-Normalization.git`.
This code includes Python scripts for preprocessing, model training, model evaluation, and also includes the JSON files for labeling.

References

Dagdelen, J., Dunn, A., Lee, S. et al. Structured information extraction from scientific text with large language models. Nat Commun 15, 1418 (2024).
https://doi.org/10.1038/s41467-024-45563-x

Trewartha, A. et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns 3, 100488 (2022).
https://doi.org/10.1016%2Fj.patter.2022.100488

Zhao, X., Greenberg, J., An, Y. & Hu, X. T. Fine-tuning BERT model for materials named entity recognition. In: 2021 IEEE International Conference on Big Data (Big Data) (IEEE, 2021).
https://doi.org/10.1109/bigdata52589.2021.9671697

Weston, L. et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. J. Chem. Inform. Modeling 59, 3692–3702 (2019).
https://doi.org/10.1021%2Facs.jcim.9b00470

Acknowledgements

This research was conducted in collaboration with I-Medata, AI and Data Center at Tel-Aviv Sourasky Medical Center (Ichilov), and The Hebrew University of Jerusalem.