L-Università ta' Malta
Faculty of Information &
Communication Technology

B.Sc. IT (Hons.)
Artificial Intelligence

Remedi: A Medical Information Extraction System

Valerija Holomjova
Supervisor: Dr. Charlie Abela
Co-Supervisor: Mr. Dylan Seychell

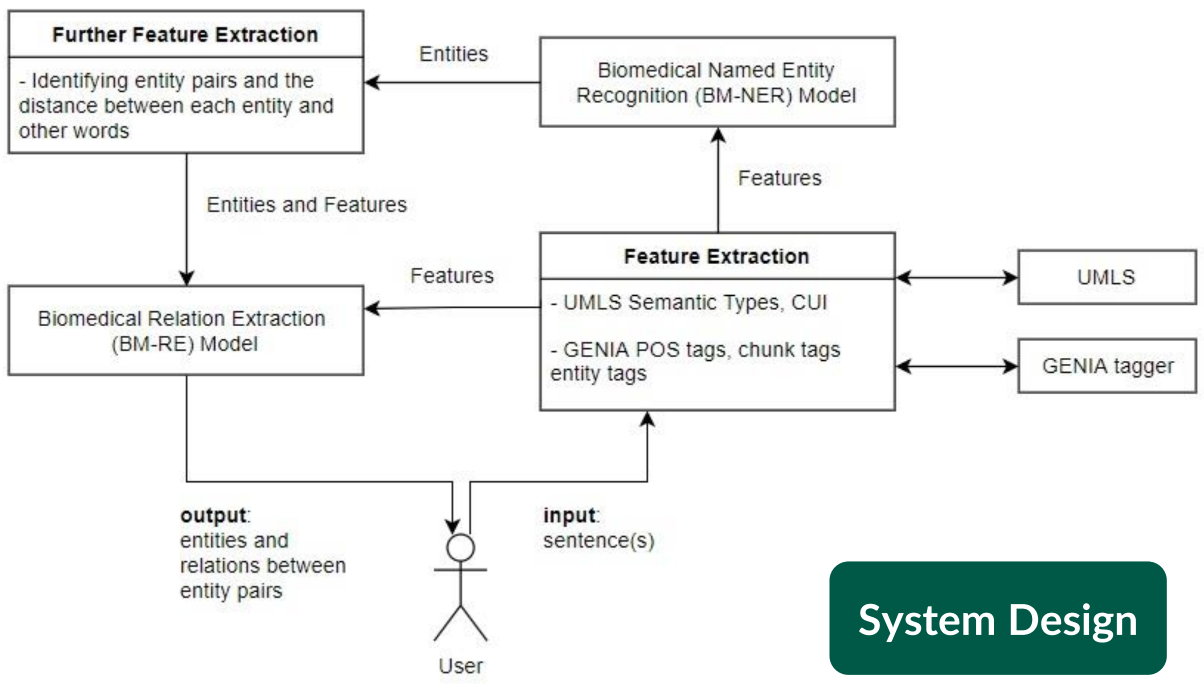
INTRODUCTION

Medical knowledge can often be found in narrative form such as medical textbooks, clinical records or online published medical reports. Numerous medical reports are being published online daily on websites containing information on advancements in the medical industry. Automated information extraction tools have been used to infer new structured knowledge from such sources with the scope of supporting both researchers and medical students. Biomedical Named-Entity Recognition (BM-NER) focuses on the recognition of medical entities such as genes, proteins or disease names. Biomedical Relation Extraction (BM-RE) is the task of extracting semantic relations between two or more medical entities. In our study, we present a system called Remedi, that is targeted at identifying medical problem and treatment entities (BM-NER) and extracting medical problem-treatment and medical problem-medical problem relations (BM-RE) from unstructured biomedical text.

AIM

- To research and develop a BM-NER tool that is able to locate and distinguish between problem and treatment entities in unstructured biomedical data.
- To research and develop a BM-RE solution that identifies relations between medical problem and treatment entities.
- To compare the model's performance with additional features [1, 2].
- To find the best model parameters through hyperparameter tuning.
- To compare the performance of DL models with baseline ML models.

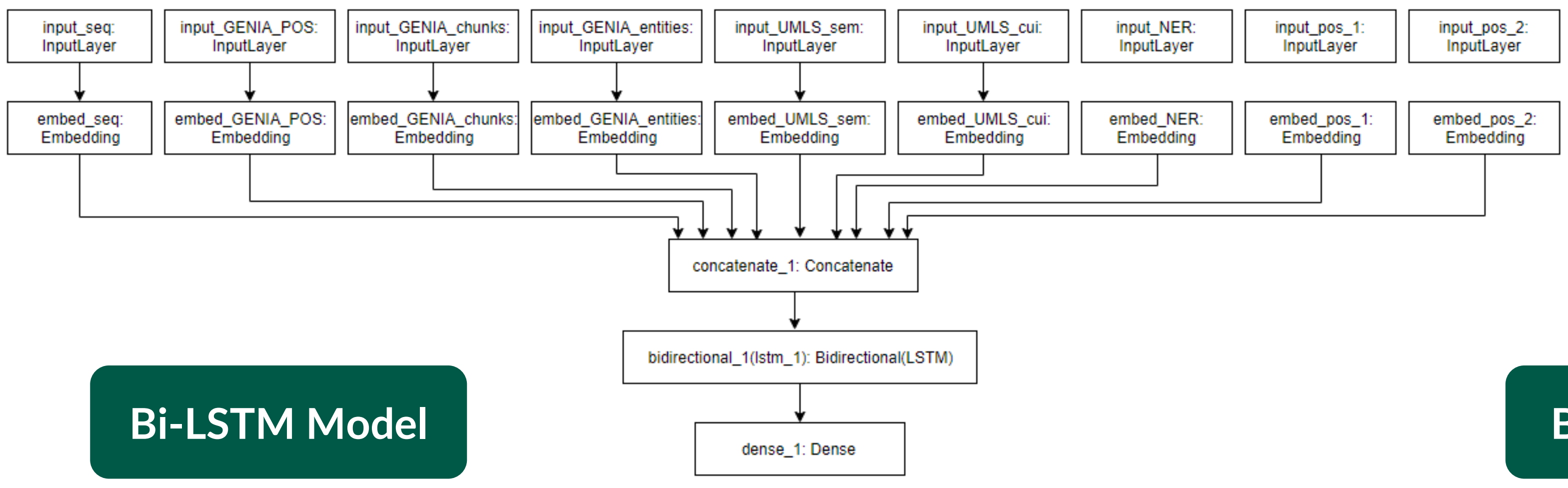
ARCHITECTURE DESIGN


Further Feature Extraction
- Identifying entity pairs and the distance between each entity and other words
Entities
Biomedical Named Entity Recognition (BM-NER) Model
Entities and Features
Features
Features
Feature Extraction
- UMLS Semantic Types, CUI
- GENIA POS tags, chunk tags entity tags
UMLS
GENIA tagger
Biomedical Relation Extraction (BM-RE) Model
output: entities and relations between entity pairs
input: sentence(s)
User

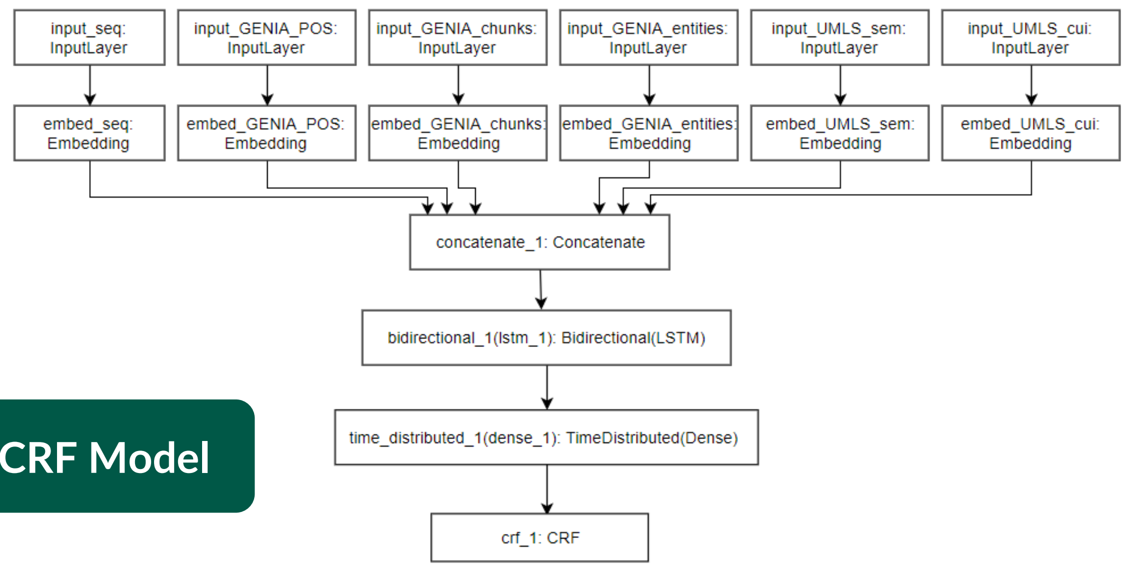
System Design

METHODOLOGY

- ❑ The Bi-LSTM-CRF and Bi-LSTM models created for the BM-NER and BM-RE components of Remedi were constructed using Keras.
- ❑ Both models were trained and evaluated using the i2b2 2010 dataset [3].
- ❑ The models use features from the UMLS [1] and the GENIA tagger [2].
- ❑ The BM-RE component also uses labelled entities from BM-NER and the distances between entity pairs and other words in sentences as features.
- ❑ The final system outputs all detected entities from given sentences and the relations between any found entity pairs.


input_seq: InputLayer
input_GENIA_POS: InputLayer
input_GENIA_chunks: InputLayer
input_GENIA_entities: InputLayer
input_UMLS_sem: InputLayer
input_UMLS_cui: InputLayer
input_NER: InputLayer
input_pos_1: InputLayer
input_pos_2: InputLayer
embed_seq: Embedding
embed_GENIA_POS: Embedding
embed_GENIA_chunks: Embedding
embed_GENIA_entities: Embedding
embed_UMLS_sem: Embedding
embed_UMLS_cui: Embedding
embed_NER: Embedding
embed_pos_1: Embedding
embed_pos_2: Embedding
concatenate_1: Concatenate
bidirectional_1(lstm_1): Bidirectional(LSTM)
dense_1: Dense

Bi-LSTM Model


input_seq: InputLayer
input_GENIA_POS: InputLayer
input_GENIA_chunks: InputLayer
input_GENIA_entities: InputLayer
input_UMLS_sem: InputLayer
input_UMLS_cui: InputLayer
embed_seq: Embedding
embed_GENIA_POS: Embedding
embed_GENIA_chunks: Embedding
embed_GENIA_entities: Embedding
embed_UMLS_sem: Embedding
embed_UMLS_cui: Embedding
concatenate_1: Concatenate
bidirectional_1(lstm_1): Bidirectional(LSTM)
time_distributed_1(dense_1): TimeDistributed(Dense)
crf_1: CRF

Bi-LSTM-CRF Model

RESULTS

- ○ Remedi's Bi-LSTM-CRF model achieved a micro F1 score of 0.846, which can be ranked amongst the top-performing models in the i2b2 2010 concept extraction challenge.
- ○ The Bi-SLTM-CRF model obtained a higher macro and micro F1 score in comparison to research that used the same model type but did not use features from the GENIA tagger and the UMLS.
- ○ The Bi-LSTM-CRF model did not outperform the baseline CRF model. This could be due to the fact that the baseline CRF model used more training features.
- ○ Remedi's Bi-LSTM achieved a micro F1 score of 0.652, and needs further improvement to reach the performance of the best models in the i2b2 2010 relation extraction challenge.
- ○ The Bi-LSTM model managed to obtain a higher micro-average precision, recall and F1 score compared to models that used the same subset of the dataset but did not leverage on the same training features.
- ○ The Bi-LSTM model was able to outperform the baseline SVM model.

CONCLUSIONS AND FUTURE WORK

In our research we presented Remedi, a system tasked with identifying medical entities in unstructured text and the relations between them. We accomplished our objectives by creating a Bi-LSTM-CRF and Bi-LSTM model for the BM-NER and BM-RE component of Remedi. The performance of the Bi-LSTM and Bi-LSTM-CRF models both improved when given features from the UMLS and the GENIA tagger, and outperformed similar models that did not use these features. For future work we would like to experiment with a larger training corpus, more training features and create a tool which uses Remedi to automatically infer and model structured knowledge from online medical articles found on sites like PubMed.

REFERENCES

1. O. Bodenreider, “The unified medical language system (UMLS): integrating biomedical terminology,” *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.
2. Y. Tsuruoka, “GENIA tagger: Part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text,” *Available at: www-tsujii. is. su-tokyo. ac. jp/GENIA/tagger*, 2006.
3. O. Uzuner, B. R. South, S. Shen, and S. L. DuVall, ¨ “2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.