Encoder-Attention based Automatic Term Recognition

Sampritha H. Manjunath, John P. McCrae
Insight SFI Research Centre for Data Analytics
Data Science Institute, National University of Ireland Galway

sampritha.manjunath@insight-centre.org

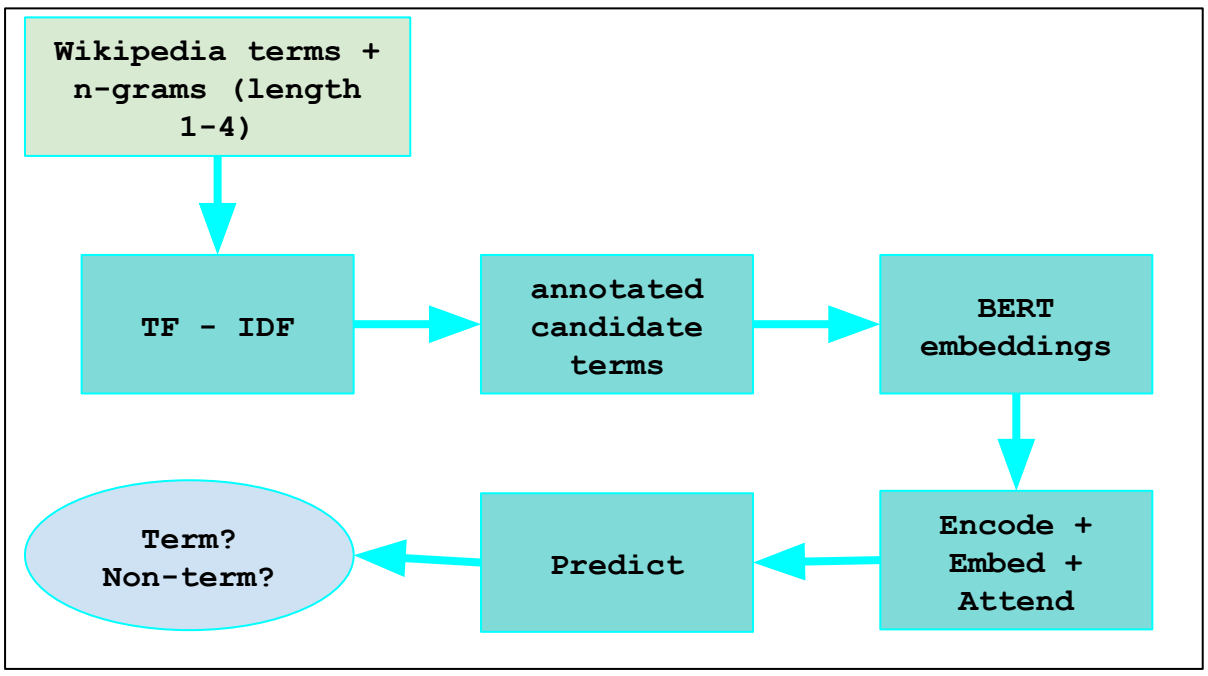
Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

A World Leading SFI Research Centre

Science Foundation Ireland sfi For what's next

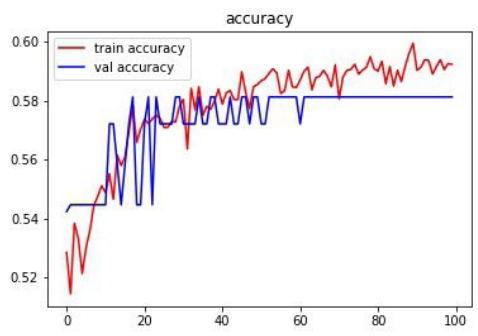
Introduction

- Creation of a binary classifier using deep learning approach to identify terms and non-terms.
- Wikipedia titles are used as positive examples and weak n-grams (of length 1-4) are used as negative examples for Automatic term recognition
- Transformer's BERT is used for word embedding.
- Encode, Embed, Attend and Predict (EEAP) approach is used to design the neural network for classification

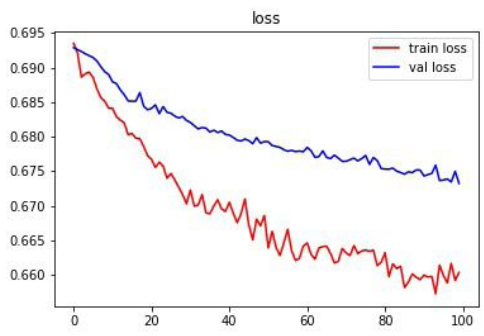


Methodology

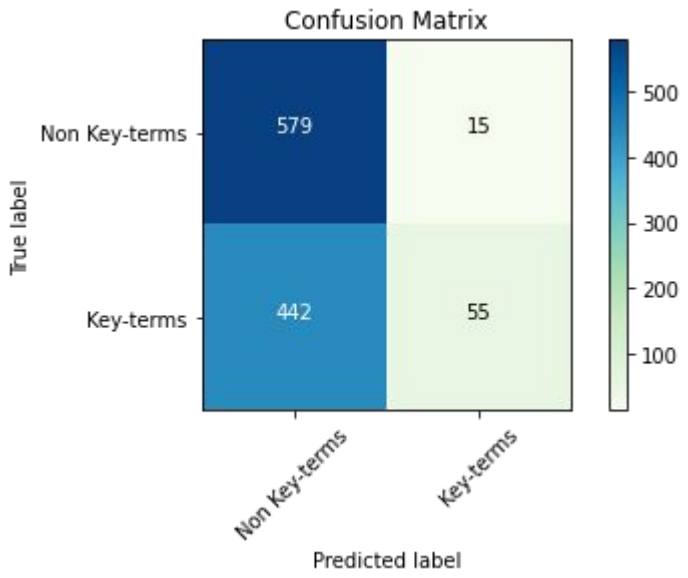
- The project can mainly be divided into four major tasks: data pre-processing, candidate term extraction, training data preparation, and modeling.
- Extract n-grams (of length 1-4) and filter them based on scores obtained by TF-IDF. This gives us the candidate terms.
- Train the EEAP model along with BERT embeddings to predict whether the given n-gram is a term or non-term



Model accuracy over 100 iteration



Decrease in loss over 100 iteration



Confusion matrix for ATR classification

Result

- EA-ATR model is evaluated against the ATR4S[1] model.

Comparison - EA-ATR(A) vs ATR4S(B)					EA-ATR model	
Dataset	**A precision**	**B precision**	**A accuracy**	**B accuracy**	**F1-score**	**Recall**
GENIA	0.8045	0.7760	60%	24%	0.7460	0.6955
Krapivin	0.6345	0.4279	62%	42%	0.7612	0.9511

Conclusion

- The model performs 28% better than the ATR4S [1] base model.
- This method has the potential to be used as a multilingual model as it does not require any annotations.
- The experiments are a clear example of a deep learning model being applied to NLP tasks by reducing the repetitive computational requirement for each dataset and to extract automatic terms more precisely.

[1] Nikita Astrakhantsev. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala. Language Resources and Evaluation, 52(3):853–872, 2018.

HOST INSTITUTIONS
DCU · NUI Galway Ollscoil na Gaillimhe · UCC University College Cork, Ireland · UCD Dublin

PARTNER INSTITUTIONS
Maynooth University · Tyndall · Trinity College Dublin · University of Limerick

FUNDED BY:
European Union European Regional Development Fund · Science Foundation Ireland sfi For what's next