

# Syapse at the NTCIR-16 RealMed-NLP task

Benjamin Holmes

Syapse, Inc.  
U.S.A

benjamin.holmes@syapse.com

Adam Gagorik

Syapse, Inc.  
U.S.A

adam.gagorik@syapse.com

Joshua Loving

Syapse, Inc.  
U.S.A

joshua.loving@syapse.com

Foad Green

Syapse, Inc.  
U.S.A

foad.green@syapse.com

Hu Huang

Syapse, Inc.  
U.S.A

hu.huang@syapse.com

## ABSTRACT

In this paper, we present our approach to subtasks 1-RR-EN, 2-RR-EN, and 3 ADE and CI of the NTCIR-16 RealMed-NLP challenge. For these challenges, the English language corpora (CR-EN and RR-EN) were used. In subtasks 1 and 2, the goal was to create an NLP system which could add tags to case reports (CR) or radiology reports (RR). In subtask 3, two applications of this system were tested: the ability to determine which RRs from a group referred to the same sample, and the ability to determine the probability that a medication caused side effects in a report. Our approach leveraged keyword extraction through a medical metathesaurus (MetaMap), sentence structuring using a SciSpacy model, and word embeddings using a trained BERT model. Using this approach, we were able to complete these three subtasks with high levels of accuracy.

## KEYWORDS

natural language processing, named entity recognition, medical report tagging

## TEAM NAME

Syapse

## SUBTASKS

Subtask 1-RR-EN (English)

Subtask 2-RR-EN (English)

Subtask 3-CR-EN (ADE)

Subtask 3-RR-EN (CI)

## 1 INTRODUCTION

The extraction of important medical information from free-text is an increasingly urgent need. In addition, advanced machine learning algorithms are being leveraged to provide better decision-making for patients [1]. These algorithms frequently rely on high-quality tags, extracted from free-text [2]. In this study, we develop a Natural Language Processing (NLP) pipeline which automatically adds customizable tags to a free-text medical document. These tags can then be extracted and used for further analysis on the document.

This system for tag addition incorporates the MetaMap semantic tagging tool [3] for keyword tagging - a system which leverages the UMLS metathesaurus for automated detection and tagging of medically relevant words. In order to properly extract not only the relevant keywords, but also the sub-pieces of the sentence which are related to the keywords, an analysis of the sentence structure is carried out using ScispaCy for sentence structuring and part-of-speech (POS) tagging [4]. This system does not rely

Syapse, May, 2022, San Francisco, California USA

B. Holmes et al.

on training examples, and so can be used essentially unmodified for subtasks 1 and 2 [12].

Once the tags are extracted, they are employed for further analysis. Subtask 3 focuses on finding Adverse Drug Events (ADEs), and on determining which reports from a group are describing the same samples. This was accomplished by using a BERT model trained on medical corpora to find similarities in extracted tags [5]. This, combined with information added to the tags in subtasks 1 and 2, allowed for these determinations to be made.

## 2 RELATED WORK

Previously, this team has worked on biomarker extraction from pathology reports using MetaMap [6]. This extraction allowed for specific biomarkers, and their testing context, to be recovered from a variety of pathology reports, despite inconsistent structuring of those reports. Indeed, the foundation of this present work can be found in this system.

MetaMap has seen use in other NLP applications for medical text, frequently for keyword extraction [7]. This is what the unmodified MetaMap algorithm accomplishes - finding specific words with relevant semantic types as defined in a medical metathesaurus.

BERT models have been used in the prediction of ADEs, leveraging custom-trained embeddings to detect these events, though much of the work has focused on their detection through social media [8][9].

ScispaCy has further been used in medical NLP, as a method of tokenizing, structuring, and extracting medically relevant sections of text corpora[10].

While these have been used in NLP applications on medical texts individually, this is the first time they are combined in a single pipeline such as the one we present here.

## 3 METHODS

First, we will address the methods used in completing subtasks 1 and 2 RR-EN. Note: because of the way our pipeline was designed, no training examples were needed, thus the procedure for tagging was identical in subtasks 1 and 2.

The first step in this pipeline is the use of MetaMap to extract relevant terms. MetaMap processes a text, and assigns attributes to the words in it.

```

"CandidateScore": "-1000",
"CandidateCUI": "C0439178",
"CandidateMatched": "% Positive",
"CandidatePreferred": "percent positive cells",
"MatchedWords": ["positive"],
"SemTypes": ["qnco"],
"MatchMaps": [
  {
    "TextMatchStart": "1",
    "TextMatchEnd": "1",
    "ConcMatchStart": "1",
    "ConcMatchEnd": "1",
    "LexVariation": "0"
  }
],
"IsHead": "yes",
"IsOverMatch": "no",
"Sources": ["CHV", "MTH", "NCI", "NCI_CDISC", "NLMSubSyn", "SNOMEDCT_US"],
"ConceptPis": [
  {
    "StartPos": "0",
    "Length": "8"
  }
],
"Status": "0",
"Negated": "0"

```

Fig. 1: An example of MetaMap features assigned to the text “% Positive”. The semantic type here is “qnco”, for “quantitative concept”

While there are many such attributes, including part of speech tagging, relative position in the document, and negation status, the attribute of primary interest here is the word’s semantic type. These semantic types are drawn from a list of UMLS-defined types, including anatomical feature, dysfunction or syndrome, quantitative concept.

These types map onto the tag types indicated by NTCIR. By testing our system on outside data sources (medical reports obtained freely from Kaggle) [11], we were able to associate semantic types with specific NTCIR-related tags.

An initial scan through the document with MetaMap allowed us to find these tag-relevant words.

Notably, some tokens in the document were not tagged by MetaMap, and had to be found through pattern matching. These included tokens dealing with time (e.g., “11:00” receives no MetaMap tags), and some terms relating to quantities (e.g., MetaMap does not recognize the token  $\mu\text{L}$  for microliter). These tags were manually added to a list, and extracted through regex operations.

Finding keywords alone was not enough to complete the tagging: there were supporting words that also needed to be applied. The next step of the pipeline used ScispaCy

## Syapse at the NTCIR-16 RealMed-NLP task

to determine the structure of sentences and extract the correct substructures within the sentence.

ScispaCy uses a spaCy model trained on biomedical data, to ensure that it recognizes and correctly tags relationships between tokens in medical texts. Once so tagged, it constructs sentence trees. These trees can be viewed with displaCy, as shown in figure 2.

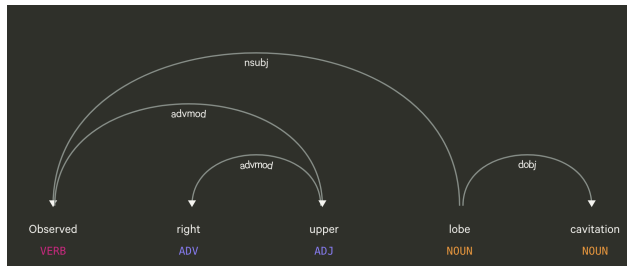


Fig. 2: An example of a sentence tree extracted by spaCy. The ‘root’ of the tree and relationships between this root and its’ supporting words are displayed graphically.

By extracting these trees, we were able to determine which words in the sentence modified the keywords identified in the MetaMap step.

Extraction of these sentence fragments was performed in two methods: firstly, by finding words which had their root in the MetaMap extracted keyword, and secondly, by matching parts of speech (POS) in these phrases. By fine-tuning which phrases were extracted using POSh, the kinds of phrases typical of certain tags (for example, adjectives modifying a noun for anatomy) could be identified.

Furthermore, a ‘ranking system’ was used to determine tag ownership if there were multiple tags which could be associated with a particular sentence fragment. This ranking system was informed by NTCIR guidelines, which specified that, for instance, disease and time generally take precedence over other tags.

Because this system made use of a pre-trained Metathesaurus from MetaMap, and a pre-trained spaCy model from scispaCy, performing subtask 2, in which no training examples were provided, was not significantly

Syapse, May, 2022, San Francisco, California USA

different from subtask 1, in which 100 training examples were provided. These training examples served as just more sentences to be processed by our pipeline.

Some tags required other attributes to be assigned. These frequently included the ‘status’ of the tag: if they had been planned, canceled, or executed. Deciding these attributes was contextually possible with the use of POS tagging performed by scispaCy, and negation tracking using MetaMap. One feature spaCy models provide is determining the tense of the words being discussed. For instance, medications and procedures which had yet to happen would be in future tense, while those executed would be in past tense. Negation tracking through MetaMap was helpful in finding scheduled remedies or medications that had been planned, but canceled.

Once the text was correctly tagged, these tags were extracted and used for further processing of the text.

Subtask 3’s Adverse Drug Event (ADE) task provided the teams with specific tags, either of a medication, or a side effect. They were asked to determine how likely it was that the tag caused a side effect (if a medicine), or was caused by a medicine (if a symptom).

This task was first passed through a simple filter. Medications that were canceled, or scheduled but not yet executed, had no chance of causing side effects. Likewise, symptoms that were negated, and thus not seen, were clearly not caused by medications.

Once passed through this simple filter, the embeddings present in the medically-trained BERT model (en\_core\_sci\_scibert) were used to determine the likelihood that a medication and a symptom were related. Cosine similarity was found between the provided tag, and the other medications tagged in the document that had the ‘executed’ state (if the provided tag was a symptom), or the other symptoms tagged in the document that were ‘observed’ (if a medication).

If there was a strong match between these tags, it was considered ‘likely’ that an Adverse Drug Event had

Syapse, May, 2022, San Francisco, California USA

B. Holmes et al.

occurred. Low matches meant a low probability that an ADE had occurred.

The CI challenge for subtask 3 was even more straightforward. In this case, anatomy, disease/symptom, and change tags were extracted from all documents and compared against each other. Documents that had the highest level of similarity, as determined by the BERT embedding, between their tags were considered to be describing the same sample. One additional factor was used in scoring these samples, i.e., special preference was given to a pre-defined size of the structures in question. Tumors described as varying by five or more mm were considered unlikely to be the same image.

#### 4 EXPERIMENTS

For subtasks 1 and 2, we achieved a character-accuracy of 82%, and an accuracy for diseases and t-tests of ~65%. This result indicates that our pipeline was able to extract and label these tags from the provided texts, and to associate the correct attributes to them. In particular, the performance on subtask 2 shows that the system can perform adequately with no training data, only samples as would normally be provided to a certified tumor registrar (CTR).

For subtask 3- ADE challenge, our team achieved top-ranking performance, indicating that keyword extraction plus embedding was sufficient to correctly extract both positive and negative adverse event descriptions. The relatively lower recall for ADE at value 3 (very likely) indicates that some adverse reactions had characteristics not picked up by the algorithm, while the high precision indicates that, once evaluated, the pipeline was excellent at determining the accuracy of these events.

<b>ADEval=0</b>	<b>P</b>	97.02
	<b>R</b>	97.63
	<b>F</b>	97.32
<b>ADEval=1</b>	<b>P</b>	30.00
	<b>R</b>	31.58
	<b>F</b>	30.77
<b>ADEval=2</b>	<b>P</b>	
	<b>R</b>	
	<b>F</b>	
<b>ADEval=3</b>	<b>P</b>	100.00
	<b>R</b>	26.32
	<b>F</b>	41.67
<b>Report-level</b>	<b>P</b>	50.00
	<b>R</b>	88.89
	<b>F</b>	64.00

For the the CI challenge in part 3, we achieved an accuracy score of ~0.8. This suggests that, overall, the system was able to locate samples that corresponded to the same sample.

These results indicate that the pipeline was able to handle reports with no training examples, to generalize to multiple report types, and to predict relationships between terms in a report. It did struggle with spatial relationships, identifying similar anatomical areas as being less related than they were.

#### 5 CONCLUSIONS

Our team was very pleased with the performance of this pipeline. While developed under time constraints, it managed to perform on par with other teams, and was able to operate with no training data, a feat not attempted by the others.

Additionally, this pipeline was designed to be modular such that it can accommodate a number of different tagging

Syapse at the NTCIR-16 RealMed-NLP task

Syapse, May, 2022, San Francisco, California USA

systems, and can operate on text with a wide variety of formattings.

A number of improvements are suggested for this pipeline in future work: training custom BERT models on specific corpora rather than using off-the-shelf models, combined with more sophisticated extractions from scispaCy models should yield more precise results.

Additionally, performing initial keyword extraction with systems other than MetaMap, such as MedCat or cTakes could make for interesting comparisons.

Overall, this NTCIR challenge was an excellent prompt to extend our NLP system, and to incorporate other technologies into its pipeline.

## REFERENCES

- [1] Jain K, Prajapati V. NLP/Deep Learning Techniques in Healthcare for Decision Making. *Prim Health Care*. 2021;(3):4.
- [2] Yadav V, Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *ArXiv191011470 Cs*. Published online October 24, 2019. Accessed March 2, 2022. <http://arxiv.org/abs/1910.11470>
- [3] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. Published online 2001:17-21.
- [4] Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proc 18th BioNLP Workshop Shar Task*. Published online 2019:319-327. doi:10.18653/v1/W19-5034
- [5] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs*. Published online May 24, 2019. Accessed March 2, 2022. <http://arxiv.org/abs/1810.04805>
- [6] Customizable Natural Language Processing Biomarker Extraction Tool | JCO Clinical Cancer Informatics. Accessed March 2, 2022. <https://ascopubs.org/doi/full/10.1200/CCI.21.00017>
- [7] Chiaramello E, Pinciroli F, Bonalumi A, Caroli A, Tognola G. Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *J Biomed Inform*. 2016;63:22-32. doi:10.1016/j.jbi.2016.07.017
- [8] Portelli B, Passabi D, Lenzi E, Serra G, Santus E, Chersoni E. Improving Adverse Drug Event Extraction with SpanBERT on Different Text Typologies. *ArXiv210508882 Cs*. Published online May 18, 2021. Accessed March 2, 2022. <http://arxiv.org/abs/2105.08882>
- [9] Hussain S, Afzal H, Saeed R, Iltaf N, Umair MY. Pharmacovigilance with Transformers: A Framework to Detect Adverse Drug Reactions Using BERT Fine-Tuned with FARM. *Comput Math Methods Med*. 2021;2021:e5589829. doi:10.1155/2021/5589829
- [10] MedLinker: Medical Entity Linking with Neural Representations and Dictionary Matching | SpringerLink. Accessed March 2, 2022. [https://link.springer.com/chapter/10.1007/978-3-030-45442-5\\_29](https://link.springer.com/chapter/10.1007/978-3-030-45442-5_29)
- [11] Electronic medical records of patients in different hospitals | Data Science and Machine Learning. Accessed March 2, 2022. <https://www.kaggle.com/general/a>
- [12] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, Eiji Aramaki. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task, In *Proc. of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.