CRF-based recognition of invasive fungal infection concepts in CHIFIR clinical reports

Yang Meng
Department of Computer Science,
University of Chicago
ymeng3@uchicago.edu

Vlada Rozova
Centre for Digital Transformation of Health,
University of Melbourne
vlada.rozova@unimelb.edu.au

Karin Verspoor
School of Computing Technologies, RMIT University
karin.verspoor@rmit.edu.au

Abstract

Named entity recognition (NER) in clinical documentation is often hindered by the use of highly specialised terminology, variation in language used to express medical findings and general scarcity of high-quality data available for training. This short paper compares a Conditional Random Fields model to the previously established dictionary-based approach and evaluates its ability to extract information from a small corpus of annotated pathology reports. The results suggest that including token descriptors as well as contextual features significantly improves precision on several concept categories while maintaining the same level of recall.

1 Introduction

Invasive fungal infections (IFIs) are a significant medical concern, particularly, among immunocompromised individuals. These infections, caused by fungal pathogens that breach the body's primary barriers and infiltrate deeper tissues or disseminate through the bloodstream, can lead to severe morbidity and heightened mortality rates. Early detection and appropriate antifungal treatment are paramount, but they may be difficult to identify in clinical populations (Even et al., 2011).

To support IFI surveillance, Rozova et al. (2023b) sought to establish an automated system to identify markers of IFI in cytology and histopathology reports. The authors introduced a corpus called CHIFIR (Rozova et al., 2023a), the Cytology and Histopathology Invasive Fungal Infection Reports, to support the development and evaluation of NLP methods for concept recognition of clinical concepts relevant to IFIs.They constructed an annotation framework to detect specific terms directly indicative of a confirmed IFI diagnosis. Central to their methods was a dictionary-based approach, which relied on exact term matches in texts.

However, the dictionary-based approach has several limitations:

- Lexical variation: the same entity can be described in different ways which complicates the task of exact matching. As an illustration, while "lung" is categorized as *Positive*, its synonym "pulmonary" is not recognized by the dictionary.
- Context is paramount: a term can convey different meanings based on its surrounding text and where in the report it is located. For instance, while "cryptococcal organism" is classified as "Fungus", the term "organism" alone may refer to bacteria, fungi, etc.

In contrast, Machine Learning (ML) algorithms, when compared with dictionary methods, present a promising alternative. These algorithms have the capability to learn the patterns of usage of relevant concepts or entities, based on consideration of the context of words.

In this work, we aim to explore the effectiveness of the ML approach by applying Conditional Random Fields (Lafferty et al., 2001) to the CHIFIR dataset and comparing its performance with the original dictionary-based solution.

The following sections will delve deeper into the methodology and outline the results of this comparison, highlighting the advantages of CRF over the dictionary approach.

2 Background

Histopathology reports are structured documents that outline findings from microscopic examination of biopsied tissue. The language used is specialized, often employing a combination of medical terminology, abbreviations, and sometimes subjective descriptions based on the pathologist's observations and interpretations. The complexity and variability of the narrative, which can differ between pathologists and institutions, make standardization difficult. Subtle nuances and contextually driven interpretations are pivotal in histopathology, making it challenging for algorithms to consistently interpret and draw accurate conclusions. Moreover, the occasional use of ambiguous or equivocal terms to describe uncertain or borderline findings can further complicate machine interpretation.

Extracting relevant concepts from clinical reports is part of a broader field of information extraction (IE). Several rule-based systems and dictionary-based entity recognition tools have been proposed offering more flexibility to combat the inherent variability in language. For instance, Funk et al. (2014) compares a ConceptMapper (Tanenblatt et al., 2010) based system with MetaMap (Aronson and Lang, 2010). While these methods offer reliability and precision, they still lack the ability to make context-specific interpretations.

In this paper, we will focus on applying Conditional Random Fields (CRFs) (Sutton et al., 2012; Sha and Pereira, 2003; Lafferty et al., 2001), which have been successfully applied to the related task of named entity recognition. CRFs are a class of statistical modeling methods and are particularly well-suited for sequence labeling tasks. CRFs consider the entire sequence, allowing for a more comprehensive contextual understanding. Additionally, CRFs are capable of ingesting a diverse set of features which can be helpful in dealing with linguistic nuances and inconsistencies across different reports. The model's flexibility enables it to effectively handle ambiguities in clinical narratives. One successful example of CRF implementation for biomedical entity recognition is BANNER (Leaman and Gonzalez, 2008).

It is worth noting that the rapid advancements in deep learning have led to the emergence of more sophisticated models, such as LSTMs and Transformer-based architectures. Such models can automatically extract features and have demonstrated superior performance across a variety of NLP tasks (Chiu and Nichols, 2016) (Santos et al., 2015). Recent literature has suggested that the use of contextualised lexical representations (e.g. in BERT (Vaswani et al., 2017)) as well as the ability to capture long-range dependencies and semantic relationships in text (Lample et al., 2016) may be particularly useful in the complex and nuanced domain of histopathology reports. However, such models might not be effective in learning IFI-specific terms because of the small and specialised nature of the CHIFIR dataset.

3 Methods

3.1 Dataset

The dataset employed for this research is the CHIFIR corpus (Rozova et al., 2023a)[1], consisting of 283 cytology and histopathology reports pertaining to 201 patients.

A characteristic feature of the cytology and histopathology reports is their extended textual format, with CHIFIR reports having an average character count of 1,384. These reports have a semi-structured layout, with headers delineating various segments for clinical annotations, macroscopic assessments, microscopic evaluations, and conclusive diagnoses.

3.2 Preparation of dataset

In this study, partitioning into development (n=230) and test (n=53) sets was replicated exactly from the original study. To ensure the results are comparable to the original study, the same stratified group k-fold cross-validation with 10 splits was applied to the development set.

Using gold standard annotations, we identified known concepts in text reports and labeled them with the corresponding categories (Table 1). The class distribution of labels is displayed in Table 2. The remaining text was tokenized into individual tokens, and each token was labeled with a default `O` label.

3.3 Model & Features

We utilized CRFSuite (Lafferty et al., 2001) as an implementation for the model and a proper set of features is needed to capture the underlying patterns in the data. We expect these features should be able to generalize, i.e., correctly discriminate the entities on new samples.

We included features that offer information on how a word appears in the text (i.e., capitalization, prefixes, suffixes) and its context. We conducted an empirical evaluation to refine the feature set: we experimented with adding semantic features, such as POS tags and special characters; sentence-level position features, such as if the word is at the start or the end of a sentence; and word-level context features, such as previous word and next word. Contrary to our intuition, the inclusion of

[1] https://physionet.org/content/corpus-fungal-infections/1.0.0/

Concept	Description
ClinicalQuery	Clinical query of IFI indicates the presence of an IFI.
FungalDescriptor	Descriptor for the presence of fungal organism.
Fungus	Mentions of specific fungal organisms.
Invasiveness	Descriptors for the depth and degree of fungal invasion into tissues.
Stain	Histological stains used to visualize fungal elements.
SampleType	Specification of the sampled organ, site, or tissue source.
Positive	Affirmative expression.
Equivocal	Expression of uncertainty.
Negative	Negating expression.

Table 1: List of concepts related to the IFI diagnosis.

Concept	Total occurrences	#reports with at least one occurence	#unique phrases	Lexical diversity
ClinicalQuery	65	53	36	0.55
FungalDescriptor	282	128	67	0.24
Fungus	106	60	15	0.14
Invasiveness	37	12	25	0.68
Stain	172	100	13	0.08
SampleType	198	179	55	0.28
Positive	118	42	37	9.31
Equivocal	7	5	5	0.71
Negative	152	104	11	0.07

Table 2: Summary statistics for the IFI-related concepts in the CHIFIR dataset.

those features either did not improve or worsened the performance of the model. The final list of included features appears in Table 3.

3.4 Experimental Framework

We tokenize each report and extract relevant features as described above. To tune hyperparameters and refine the feature set, we used cross-validation whereby within each fold, a CRF model is initialized with ‘lbfgs’ algorithm and a maximum iteration of 100. The final model with hyperparameters c1=0.01 and c2=0.01 was trained on the entire training dataset to generate predictions on the test set.

For evaluation, we used full-term identification. We calculated the number of true positive, false positive, and false negative concepts in each report by comparing the predictions to the gold standard annotations. For each concept category, we summarize model performance using precision and recall, and record incorrectly identified concepts for error analysis.

4 Results

4.1 Overview

Overall, the CRF approach outperformed the dictionary-based approach utilized in the original paper (Rozova et al., 2023b). Table 4 shows a significantly higher precision in detecting categories *FungalDescriptor*, *SampleType*, *Positive*, and *Negative*. For other concept categories, the CRF model had on average higher precision although the difference was not statistically significant. Table 5 shows that recall is on average comparable to that of the dictionary-based approach. Table 6 summarises the performance as F1 score showing significant improvement in categories *SampleType*, *Positive*, *Equivocal*, and *Negative*.

4.2 Strengths

First, let us consider the challenge of lexical variation. The ability of the dictionary-based approach to generalize is limited; to make a correct prediction a concept has to appear in the same form as in the training sample. For our CRF model, we found

Feature	Description
`word`	The word itself.
`start_pos and end_pos`	The start and end position of the word.
`is_capitalized`	Checks if the first letter is capitalized.
`is_all_caps and is_all_lower`	Check for casing details.
`capitals_inside`	Checks if there are capital letters inside the word.
`prefix and suffix`	Use the 3 prefix and 3 suffix characters of each word as context.
`has_hyphen`	Whether the word has hyphens.
`is_numeric`	Whether the word has numeric.

Table 3: List of features.

Concept	Precision CV Dict	Precision CV CRF	Precision TEST Dict	Precision TEST CRF
ClinicalQuery	0.92 (±0.13)	0.83 (±0.20)	1.00	1.00
FungalDescriptor	0.75 (±0.10)	0.92 (±0.05)	0.68	0.98
Fungus	0.82 (±0.30)	0.95 (±0.07)	0.88	0.94
Invasiveness	0.45 (±0.41)	0.69 (±0.41)	0.33	1.00
Stain	0.94 (±0.05)	0.97 (±0.05)	1.00	0.97
SampleType	0.15 (±0.03)	0.92 (±0.08)	0.14	1.00
Positive	0.04 (±0.02)	0.82 (±0.16)	0.03	1.00
Equivocal	0.01 (±0.02)	1.00 (±NaN)	0.00	0.00
Negative	0.14 (±0.04)	0.97 (±0.05)	0.15	1.00

Table 4: Comparison of dictionary and CRF approach precision during cross-validation and on unseen test data.

Concept	Recall CV Dict	Recall CV CRF	Recall TEST Dict	Recall TEST CRF
ClinicalQuery	0.53 (±0.35)	0.72 (±0.20)	0.69	1.00
FungalDescriptor	0.93 (±0.04)	0.90 (±0.05)	0.93	0.96
Fungus	0.92 (±0.15)	0.88 (±0.16)	0.94	0.94
Invasiveness	0.60 (±0.39)	0.63 (±0.30)	0.12	0.50
Stain	0.95 (±0.09)	0.98 (±0.04)	1.00	1.00
SampleType	0.86 (±0.10)	0.81 (±0.11)	0.86	0.79
Positive	0.83 (±0.17)	0.89 (±0.13)	0.73	0.95
Equivocal	0.58 (±0.50)	0.20 (±0.45)	0.00	0.00
Negative	0.98 (±0.05)	0.96 (±0.08)	0.90	1.00

Table 5: Comparison of dictionary and CRF approach recall during cross-validation and on unseen test data.

that about 82% of the correctly predicted concepts in the test set were exact matches from the training set, and the rest were variations of known concepts.

The CRF model can identify and combine parts of annotated concepts. For instance, "branching hyphae" was not present in the training set. CRF generalizes "branching" and "hyphae" by learning from two concepts in the training data, "acute angle branching" and "septate hyphae", which were annotated as *FungalDescriptor*. The suffix "cosis" was also captured as an indicator of the *Fungus* category. The model captures linguistic/capitalization/syntax variations, for instance, "duodenum" is generalized from "duodenal", and "groccot" from "Groccot". Besides, CRF demonstrated the ability to learn complex patterns: "? infection PJP" is detected based on the *FungalDescriptor* "PJP" present in the training data and the fact that a "?" followed by a *FungalDescriptor* often makes up a *ClinicalQuery*. The model captures

Concept	**F1** **CV Dict**	**F1** **CV CRF**	**F1** **TEST Dict**	**F1** **TEST CRF**
ClinicalQuery	0.68 (±0.27)	0.75 (±0.16)	0.81	1.00
FungalDescriptor	0.83 (±0.07)	0.91 (±0.03)	0.79	0.97
Fungus	0.91 (±0.09)	0.90 (±0.09)	0.91	0.94
Invasiveness	0.71 (±0.25)	0.68 (±0.27)	0.18	0.67
Stain	0.94 (±0.05)	0.97 (±0.03)	1.0	0.98
SampleType	0.26 (±0.04)	0.86 (±0.08)	0.24	0.88
Positive	0.08 (±0.03)	0.84 (±0.10)	0.05	0.97
Equivocal	0.05 (±0.03)	1.00 (±NaN)	NaN	NaN
Negative	0.24 (±0.06)	0.96 (±0.04)	0.26	1.00

Table 6: Comparison of dictionary and CRF approach F1 during cross-validation and on unseen test data.

the intuition that certain labels are more likely to appear after certain other labels. Lastly, phrases not present in the training data, such as "punch biopsies", "pericardium", and "abdomen" were correctly predicted, showing that the model can make inferences based on relevant contexts.

Secondly, the model did a generally good job of addressing ambiguity in the medical text. Words such as "organism" and "capsule" were consistently overdetected when using the dictionary-based approach, resulting in a high false-positive rate. The CRF model has correctly picked out the relevant mentions considering their context.

4.3 Weaknesses

In general, the detection of concepts belonging to *SampleType* and *Invasiveness* categories showed to be the most challenging, making up 45% and 17% of the total error cases, respectively. The errors were largely due to the relatively modest size of the training data, high lexical diversity and fewer occurrences in the dataset.

The modest recall characteristic of the *Invasiveness* category is likely due to high lexical diversity and longer phrases consisting of multiple tokens. For example, the model failed to classify phrases "tissue invasion" and "vessel lung parenchyma infiltrated" as *Invasiveness* concepts, even though individual words "invasiveness", "vessel", and "parenchyma" were frequently occurring in the training data. It is possible that engineering a more extensive contextual feature set is required to tackle such cases.

Some words did not appear in the training data and thus the model may have never learned an appropriate representation. This can be seen in examples involving both medical terms (e.g., "ileum", "cyst") and generic English words (e.g., "back", "leg").

The features used in the model may also occasionally be misleading. For instance, the word "RUL" is misclassified as *Stain* because a common *Stain* concept "PAS" usually appears in uppercase. Thus the model may associate the upper case with that label, illustrating an example of the model giving form much more weight than context.

5 Conclusion

In conclusion, we have seen that the CRF model performes better and, in particular, is more successful in tackling the lexical diversity and variation present in the CHIFIR corpus than the previous dictionary-based method. Although the model performance still suffers from the small sample size and challenging lexical diversity cases, we demonstrated that incorporation of context through the CRF-based concept recognition model benefits development of clinical concept recognition tools for this corpus. It would also be worth exploring and comparing this CRF-based approach with more advanced machine learning methods, which might be able to learn richer representations from data, and overcome challenges posed by the variability and linguistic nuances in histopathology texts better.

References

A.R. Aronson and F.-M. Lang. 2010. An overview of metamap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*

J.P. Chiu and E. Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association.*

C. Even, S. Bastuji-Garin, and Y. Hicheri. 2011. Impact of invasive fungal disease on the chemotherapy schedule and event-free survival in acute leukemia patients who survived fungal disease: a case-control study. *Haematologica*.

Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Bretonnel Cohen, Lawrence E Hunter, and Karin Verspoor. 2014. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15:1–29.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

G. Lample, M. Ballesteros, and S. Subramanian. 2016. Neural architectures for named entity recognition. *Proceedings of the HLT-NAACL*.

Robert Leaman and Graciela Gonzalez. 2008. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing 2008*, pages 652–663. World Scientific.

Vlada Rozova, Anna Khanina, Jasmine C. Teng, Joanne S.K. Teh, Leon J. Worth, Monica A. Slavin, Karin A. Thursky, and Karin Verspoor. 2023a. Chifir: Cytology and histopathology invasive fungal infection reports (version 1.0.0). *PhysioNet*.

Vlada Rozova, Anna Khanina, Jasmine C. Teng, Joanne S.K. Teh, Leon J. Worth, Monica A. Slavin, Karin A. Thursky, and Karin Verspoor. 2023b. Detecting evidence of invasive fungal infections in cytology and histopathology reports enriched with concept-level annotations. *Journal of Biomedical Informatics*, 139:104293.

C.N. Santos, V. Guimaraes, and R.J. Niteroi. 2015. Boosting named entity recognition with neural character embeddings. *Proceedings of NEWS2015 The Fifth Named Entities Workshop*.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Michael A Tanenblatt, Anni Coden, and Igor L Sominsky. 2010. The conceptmapper approach to named entity recognition. In *Language Resources and Evaluation Conference*, pages 546–51.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.