Clinical Named Entity Recognition in Indonesian Clinical Research Paper using Conditional Random Fields

Rofi Tulus Syifa[1], Nurmaya[1], Nova Eka Diana[1], Yurika Sandra[2], Wahyu Catur Wibowo[3], Indra Budi[3]
[1]Faculty of Technology Information, Universitas YARSI, Jakarta Pusat, Indonesia
[2]Faculty of Medicine, Universitas YARSI, Jakarta Pusat, Indonesia
[3]Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
{rofituluss2@gmail.com, nurmaya@yarsi.ac.id*}

Abstract

Clinical Research Papers, written in Indonesian Language, contain novelty in clinical knowledge that discusses about patient sign and symptoms, diagnosis, and medical procedures. To assist the readers quickly gain insight into the recent knowledge, it is necessary to extract the entities in Clinical Research Papers. In this study, we have built a Clinical Named Entity Recognition (CNER) model for Indonesian Language using Conditional Random Fields (CRF) with contextual, prefix & suffix, orthographic, stop word, and POS-Tagging features. Using the Indonesian Clinical Corpus annotated in Inside-Outside-Begin (IOB2) format as a data set, we extract seven clinical entities including body part, sign and symptom, disease, physical examination, supporting examination, pharmacological and non-pharmacological treatment. The experiment result shows that our model achieves the best performance in extracting multiple entities with 3 features (contextual, prefix & suffix, and ortographic) with precision of 88%, recall of 73%, and f-measure of 79%.

Objective

1. Building an Indonesian Clinical Corpus generated from 29 articles of Jurnal Penyakit Dalam Indonesia (Internal Medicine Journal of Indonesia)
2. Generating Clinical Named Entity Recognition model using CRF algorithm with contextual, prefix & suffix, orthographic, stop word, and POS-Tagging features

Material

1. An Indonesian Clinical Corpus, with the total amount of 49444 words annotated in IOB2 format by 11 fourth year students of Faculty of Medicine that are supervised by 1 faculty member of Medicine Faculty, are the dataset used by CRF to generating a Clinical Named Entity Recognition (CNER) model.
2. The data set are divided into training and testing data. The ratio of training and testing is a part of experiments to create the best CNER model.

Method

There are two main parts in building CNER model as follows : creating an Indonesian Clinical Corpus and implementing CRF with five extract features such as contextual, prefix & suffix, orthographic, stop word, and POS-Tagging features

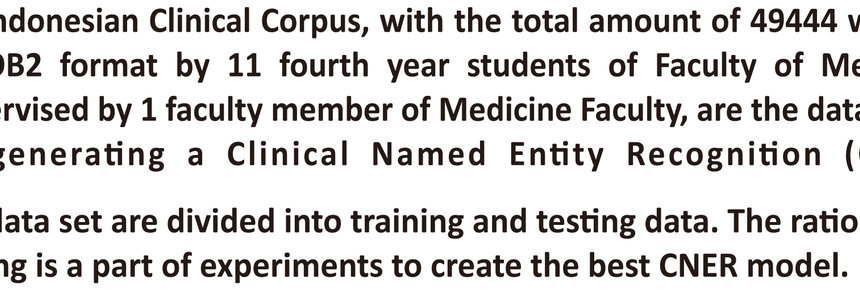
A. Creating an Indonesian Clinical Corpus

There are four steps for making the corpus: Collecting the articles, Tokenization, IOB2 Manual Tagging and POS-Tagging.

B. Implementing CRF algorithm with Five Extract Features

There are three steps including triming a sentence into a sequence of words, choosing the features, training the CRF

THE WORKFLOW OF THE SYSTEM IS GIVEN BELOW

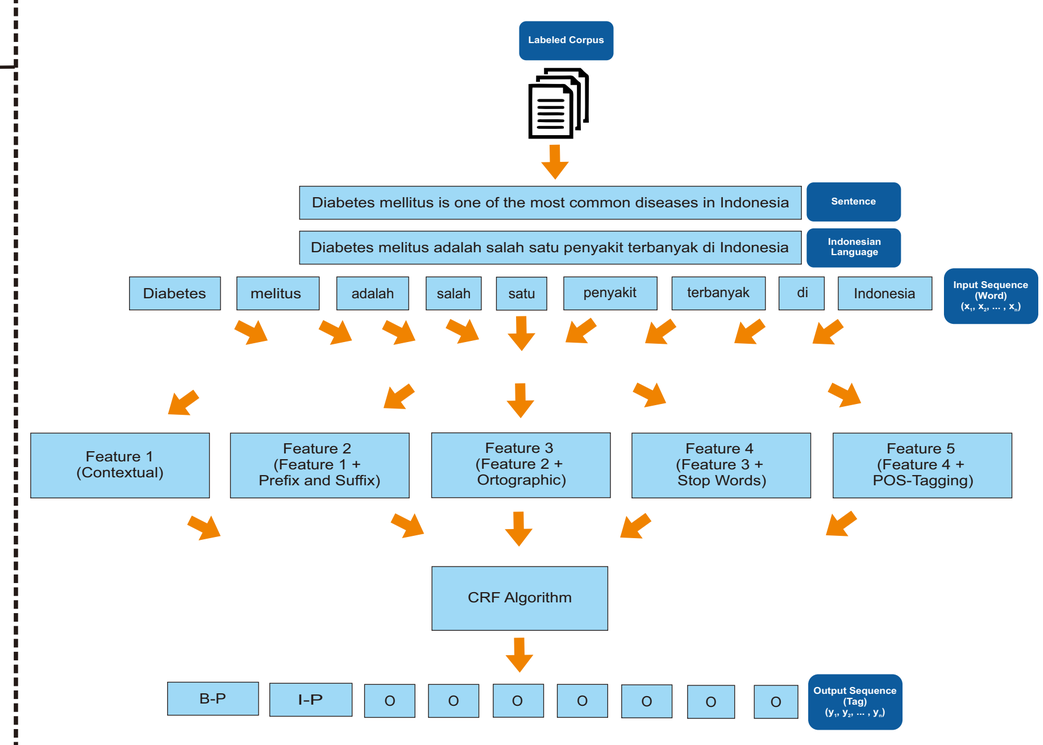


Data	Sentences	Words
Corpus	2.051	49.444

Entities	Organ (Body Part)	Sign & Symptom	Pemeriksaan Fisik (Physical Examination)	Pemeriksaan Penunjang (Supporting Examination)	Penyakit (Disease)	Terapi Farmakologi (Pharmacological Treatment)	Terapi Non-Farmakologi (Non-Pharmacological Treatment)
Format IOB2	B-ORG	B-SS	B-PF	B-PP	B-P	B-TF	B-TNF
	I-ORG	I-SS	I-PF	I-PP	I-P	I-TF	I-TNF

Entities	B-ORG	B-SS	B-P	B-PF	B-PP	B-TF	B-TNF	I-ORG	I-SS	I-P	I-PF	I-PP	I-TF	I-TNF	O
Total	185	382	1508	27	622	1067	129	100	679	907	80	708	558	102	42390
Percent	0.37%	0.77%	3.05%	0.05%	1.26%	2.16%	0.26%	0.20%	1.37%	1.83%	0.16%	1.43%	1.13%	0.21%	85.7%

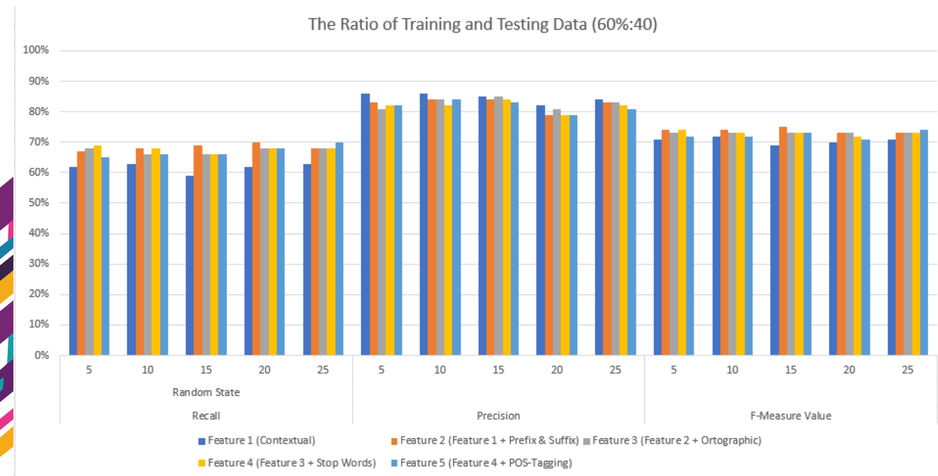
B



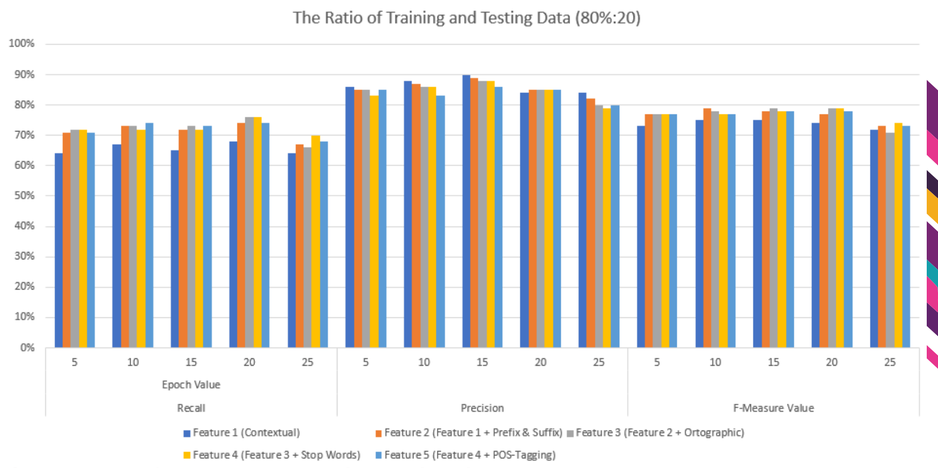
Result

The experiments conducted in this study are comparing the ratio of training and testing data with random state value and features, comparing the ratio of training and testing data with epoch value. We choose the best CNER Model based on the recall, precision, and F-Measure Value.

A. The ratio of Training and Testing Data (60%:40%) with features and random states



B. The ratio of Training and Testing Data (80%:20%) with features and random states



C. The ratio of Training and Testing Data with epoch values

Ratio	Features	Epoch	Recall	Precision	F-Measure Value
60%:40%	Feature 1 (Contextual)	50	61%	84%	70%
		100	62%	84%	71%
		150	62%	85%	71%
	Feature 2 (Feature 1 + Prefix & Suffix)	50	68%	82%	73%
		100	68%	81%	73%
		150	68%	82%	73%
	Feature 3 (Feature 2 + Ortographic)	50	67%	83%	73%
		100	67%	81%	73%
		150	67%	81%	73%
	Feature 4 (Feature 3 + Stop Words)	50	67%	81%	73%
		100	68%	81%	73%
		150	68%	80%	73%
	Feature 5 (Feature 4 + POS-Tagging)	50	67%	81%	72%
		100	67%	80%	72%
		150	67%	80%	72%
80%:20%	Feature 1 (Contextual)	50	65%	86%	73%
		100	65%	86%	73%
		150	66%	85%	74%
	Feature 2 (Feature 1 + Prefix & Suffix)	50	71%	85%	76%
		100	73%	85%	77%
		150	73%	84%	77%
	Feature 3 (Feature 2 + Ortographic)	50	72%	84%	76%
		100	73%	83%	77%
		150	72%	83%	76%
	Feature 4 (Feature 3 + Stop Words)	50	72%	84%	76%
		100	72%	82%	76%
		150	72%	82%	76%
	Feature 5 (Feature 4 + POS-Tagging)	50	72%	83%	76%
		100	71%	82%	76%
		150	72%	83%	76%

A

Conclusion

Clinical Name Entity Recognition model for Indonesian Language reaches its best performance using CRF algorithm combining with 3 features (contextual, prefix & suffix, and ortographic), the ratio of training & testing data is 80:20, 15 and 50 for random state value and epoch value. The model has precision of 88% ,recall of 73%, F-1 measure of 79%.