

# Finding Questions in Healthcare Forum Posts using Sequence Labeling Approach

Adrianus Saga Ekakristi, Rahmad Mahendra, and Mirna Adriani

Faculty of Computer Science, Universitas Indonesia  
Depok, 16424 West Java, Indonesia  
`adrianus.saga@ui.ac.id`, `rahmad.mahendra@cs.ui.ac.id`

**Abstract.** Complex medical question answering system in medical domain receives a question in form of long text that need to be decomposed before further processing. This research propose sequence labeling approach to decompose that complex question. Two main tasks in segmenting complex question sentence are detecting sentence boundary with its type, and recognizing word that could be ignored in sentence. The proposed sequence labeling method achieve 0.83 F1 score in detecting beginning sentence boundary and 0.93 F1 score when determining sentence type. In recognizing word that could be ignored in sentence, the proposed sequence labeling method achieve 0.90 F1 score.

**Keywords:** Complex Question Decomposition, Medical Question Answering, Natural Language Processing, Sequence Labeling, Chunking

## 1 Introduction

Common factoid question answering (QA) system receive a one simple question that could be answered with simple fact. This architecture of factoid QA cannot answer health-related question from society in general. When consulting with a doctor, people tend to describe their symptom in multiple sentences and then ask one or more question. Most of them even say greetings or thank you. This complex question need to be decomposed into several parts, such as contextual information about patient's disease, question, or sentence that can be ignored. Consider the following example:

- umur saya 21 tahun. sudah 4 hari ini panas badan saya turun naik, sebenarnya saya sakit apa??? terima kasih. (*my age is 21 years old. for these past 4 days, my body heat was getting up and down, what is actually my disease??? thank you.*)
- Lalu bagaimana cara pengobatannya ya, dok? (*And how to cure it, doc?*)

The first example of complex question above consists of multiple sentence background information, followed by a question, and closed with a gratitude sentence. The last sentence can be ignored because it doesn't contain useful information about patient's symptom or disease. We can also observe the use of

informal grammar and structure, such as using comma at the end of the sentence. In the second example, we can see that the user use greeting word "dok" that refer to the doctor. Greeting word like this doesn't contain any useful contextual information about patient's symptom and could be ignored. A question answering system should first be able to recognize various sentences, whether a question, a description of user's symptom, or other parts that should be ignored.

This process of decomposing complex question text is a crucial step in medical complex question answering system. In this paper, we propose sequence labeling approach for decomposing complex question text. The module we build receives a consultation request text as an input and return a list of sentences with its type as an output. The data we used for training and testing is question and answer pairs from health consultation forum.

## 2 Related Work

Even though usually, rule-based method already give high accuracy in segmenting text into sentences, the use of sequence labeling approach to enhance this process has been proposed by Evang et al. [2]. It is not only proved to increase the accuracy, but also easier to maintain than hand-crafted rules and mitigate the language-specific restriction of the method. Evang et al. [2] define a set of IOB tag that represent the beginning of a sentence, beginning of token (word), inside a token, and outside the token. Conditional Random Field was chosen as the model and character-level word embedding as the feature. The proposed method was then tested in three news corpus, each has different languages. The result show higher F1 sentence score than state-of-the-art boundary detection system.

Several literature has tried to apply NLP techniques to decompose question text into atomic question along with contextual background information. Each of those atomic question could then be answered by question answering system. Roberts et al. [6] proposed a system consists of several module to decompose English consumer-health complex question. A question text, which they called request, was segmented and classified into several sentence type, which are question, ignore, and background. The question sentence would undergo several decomposition process, which are decomposing clause-level question, decomposing phrase that spans a set of decomposable items, and decomposing phrase that spans an optional item. Furthermore, Roberts et al. [6] also proposed classification for sub-class of background sentence and focus recognition. More detailed definition of each annotation is provided in Roberts et al. [5]. The proposed methods for those modules are mostly combination of rule-based candidate generator followed by rank-and-filter machine learning methods. The tokenization and sentence segmentation process are not described in detail, most likely because rule-based method already produce a good result for relatively good-grammar questions.

Sondhi et al. [7] on the other hand, proposed a method to extract medical problem and medical treatment descriptions from medical forum data. The data

itself were taken from HealthBoards<sup>1</sup> and manually annotated. Support Vector Machine and Conditional Random Field is chosen as the model and various features were proposed, such as n-gram language model, the Unified Medical Language System (UMLS)<sup>2</sup> semantic groups of words, position features, heuristic user-based features, previous sentence tag, length-based features, and various morphological features. The proposed method achieve best prediction accuracy above 75%.

In domain of Indonesian language, Hakim et al. [3] have gather a corpus consists of medical questions taken from five different health consultation website. The corpus consists of 86731 question-answer pairs which are produced by the interaction of patient and doctor in the forum. Mahendra et al. [4] on the other hand, proposed a method for decomposing Indonesian question text in the corpus. The three main components are sentence splitter, sentence classifier, and multi-questions splitter. Sentence splitter segmented the question paragraph into a list of sentence using rules, such as delimiters and heuristic rules. In sentence classifier component, Support Vector Machine model was used to classify sentences into types defined by Roberts et al. [6]. The features used are n-grams, position and length of sentence, question-specific attributes, and dictionary of symptoms and diseases. For multi-questions splitter, several strategy was proposed to accommodate different combination of sentence types in one sentence that need to be decomposed.

### 3 Proposed Method

Our question decomposition module consists of two main process, which are sentence boundary and type recognition and ignored words recognition.

#### 3.1 Data

We used the data provided by Hakim et al. [3] which is a collection of patient-doctor question-answer pair taken from five different health consultation forum. For evaluation, we randomly sample and manually annotate 200 question for sentence boundary and sentence type recognition, along with ignored word recognition. More detail specification of annotated data is shown in Table 1 and 2. To represent the class and position of token in sequence, we use the IOB tag.

#### 3.2 Part-of-Speech Tagging

Part-of-speech (POS) tagging is one of data processing steps we apply in our experiment. Given a sequence of words (token), the task is to determine a proper part-of-speech tag for each token. We use sequence labeling approach to do POS tagging, specifically linear chain Conditional Random Field (CRF) model. We use the dataset and tagset definition provided by Dinakaramani et al. [1].

<sup>1</sup> <https://www.healthboards.com/>

<sup>2</sup> <http://www.nlm.nih.gov/research/umls/>

**Table 1.** Specification of annotated sentences

Sentence Type	Number of sentence
BACKGROUND	714
IGNORE	359
QUESTION	305
Total	1378

**Table 2.** Specification of annotated tokens

Tag	Number of token
B-BACKGROUND	714
I-BACKGROUND	8752
B-IGNORE	359
I-IGNORE	928
B-QUESTION	305
I-QUESTION	2668
Total	13726

### 3.3 Sentence Boundary and Type Recognition

Our system take a sentence written by Indonesian society as an input. We have found that informal language is a common case, such as use of abbreviated word or incorrect / ambiguous sentence boundary. Thus, one of our main task is to determine the boundary of sentences given a long sequence of token. The next task is to classify the type of each sentence. The output of this process is a list of sentence (which is a list of token) with the corresponding type. We define three type of sentence, which are described below.

1. Background: sentence that show useful contextual information, but doesn't contain question.
2. Question: sentence that contain one or more clause that express a question from user.
3. Ignored: sentence that doesn't contain any useful information or question and can be ignored. Example: "Selamat pagi, dokter" (*Good morning, doctor*)

We propose a sequence labeling approach for this task. The model we used is linear chain CRF. We also propose two strategy for detecting the sentence boundary, which are illustrated in Fig. 1.

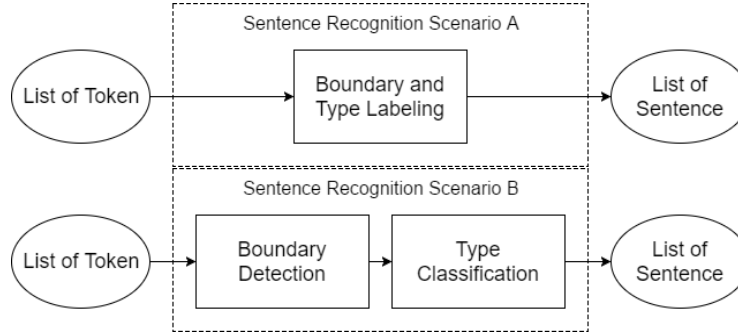
1. Strategy A

In this strategy, the prediction of boundary and type of each token is done in one sequence labeling process.

2. Strategy B

In this strategy, first, we determine which token is the boundary of sentences. Afterwards, the type of each sentence is then determined.

We use various features for our model, whether on strategy A or B. For illustration purpose, consider the following example, which will be used to give a better understanding about each proposed feature:



**Fig. 1.** Sentence Recognition Strategies

- Pagi dok, saya Flu, Apa obatnya? (*morning doc, i cougth a Flu, What is the medication?*)

The example above consists of 3 sentence. Note that there are incorrect uses of grammar and it is a very common case in our data. The first sentence ("Pagi dok,") is just a greeting which could be ignored. The second sentence ("saya Flu,") is a background sentence which contain useful symptom information. The last sentence is a question. The IOB tags for each token is shown in Table 3.

**Table 3.** Example for feature extraction simulation

Sentence 1	Token	Pagi	dok	,
	POS	NN	NN	Z
<b>Sentence 2</b>	Token	saya	<b>Flu</b>	,
	POS	PRP	NN	Z
Sentence 3	Token	Apa	obatnya	?
	POS	WH	NN	Z

In strategy A, the feature we extract for boundary and type labeling is specified below along with the value of the feature when extracted from the token "Flu" (marked in bold) in example above.

1. Features regarding the token itself:
  - (a) token itself  
value in example above: "Flu"
  - (b) token in lower form  
value in example above: "flu"
  - (c) whether the token is digit  
value in example above: *false*
  - (d) whether the first character is capital letter  
value in example above: *true*

- (e) whether all character is capital letter  
value in example above: *false*
- 2. Feature regarding the position of the token:
  - (a) position of the token  
value in example above: 5/9
  - (b) whether the token is the first token in the text  
value in example above: *false*
  - (c) whether the token is the last token in the text  
value in example above: *false*
- 3. Features regarding neighboring token:
  - (a) the token after  
value in example above: ", "
  - (b) the token before  
value in example above: "saya"
  - (c) whether the token is one of the following: period, comma, exclamation mark, question mark  
value in example above: *false*
- 4. Part-of-Speech  
value in example above: "NN" which means a noun
- 5. Website where the data from  
Every medical forum has their own unique pattern in the data. For example, the question text from Detik Health<sup>3</sup> usually append the biodata of the user which post the consultation question.
- 6. Dictionary of medical terminology  
We use medical terminologies taken from Standar Kompetensi Dokter Indonesia 2012 <sup>4</sup> (Indonesian Standard Competency of Doctor 2012).
- 7. Abbreviation dictionary  
We use dictionary of abbreviation of Indonesian word.

Below, we list the feature used for sentence boundary detection in strategy B along with the value of the feature when extracted from the token "Flu" (marked in bold) in example shown in Table 3. Most of the features used in sentence recognition in strategy A also used in this step.

- 1. Features regarding the token itself:
  - (a) token itself  
value in example above: "Flu"
  - (b) token in lower form  
value in example above: "flu"
  - (c) whether the token is digit  
value in example above: *false*
  - (d) whether the first character is capital letter  
value in example above: *true*

<sup>3</sup> <https://health.detik.com/>

<sup>4</sup> <http://pd.fk.ub.ac.id/wp-content/uploads/2014/12/SKDI-disahkan.pdf>

- (e) whether all character is capital letter  
value in example above: *false*
- 2. Feature regarding the position of the token:
  - (a) position of the token  
value in example above: 5/9
  - (b) whether the token is the first token in the text  
value in example above: *false*
  - (c) whether the token is the last token in the text  
value in example above: *false*
- 3. Features regarding neighboring token:
  - (a) the token after  
value in example above: ", "
  - (b) the token before  
value in example above: "saya"
  - (c) whether the token is one of the following: period, comma, exclamation mark, question mark  
value in example above: *false*
- 4. Part-of-Speech  
value in example above: "NN" which stands for noun
- 5. Website where the data from

To classify the type of each sentence in strategy B, the following features are used. We also include the value of each feature when extracted from second sentence (marked in bold) in the example shown in Table 3.

- 1. Unigram  
value in example above: "saya", "Flu", and ", "
- 2. Feature regarding tokens in the sentence:
  - (a) quantity of token  
value in example above: 3
  - (b) whether the sentence contain question mark  
value in example above: *false*
  - (c) whether the sentence contain question word, such as "apa" (*what*) and "bagaimana" (*how*)  
value in example above: "false"
- 3. Position of the sentence in the text  
value in example above: 2/3

### 3.4 Ignored Word Recognition

Our user often use word that doesn't contain any useful contextual information. One of the common case is greeting words, such as "dok" (a form of greeting to doctor). Given background or question sentence, the task is to identify words that could be ignored.

We propose sequence labeling approach for this task. We use linear chain CRF as our learning model. The feature we extract for our model is listed below.

1. Features regarding the token itself:
  - (a) token itself
  - (b) token in lower form
  - (c) whether the token begins with capital letter
2. Features regarding neighboring token:
  - (a) the token after
  - (b) the token before
3. The position of token in text

## 4 Result and Analysis

### 4.1 Sentence Boundary and Type Recognition

We evaluate each of our proposed strategies using 10-fold cross-validation. The result of strategy A is shown in Table 4.

**Table 4.** Performance of Sentence Recognition on Strategy A

Tag	Precision	Recall	F1 Score
B-BACKGROUND	0.83	0.71	0.77
I-BACKGROUND	0.92	0.97	0.94
B-IGNORE	0.93	0.78	0.85
I-IGNORE	0.88	0.76	0.82
B-QUESTION	0.83	0.77	0.80
I-QUESTION	0.89	0.85	0.87
Average / Total	0.91	0.91	0.91

In scenario A, we found that 71.62% of the sentences are exact match, specifically 232 out of 305 (76.06%) question sentence, 482 out of 714 (76.04%) background sentence, and 273 out of 359 (67.50%) ignored sentence.

There are also cases where the sentence are predicted partially. For example, there are predicted sentences that are cutoff in the beginning and/or end. We found 70 out of total 1378 (28.37%) sentences are partially predicted, in which 12 of them are question sentence, 47 are background sentence, and 11 are ignored sentence. The composition of those 70 partial cases is the following: 30 cases cutoff in the beginning, 36 cases cutoff in the end, and 4 cases cutoff in both beginning and end. Furthermore, the average number of token that are cutoff is 6.5 tokens in the beginning of the sentence, 21.5 in the end, and 21.5 in both the beginning and end of the sentence. Overall, the average number of cutoff token is 8.7 tokens.

Other than exact match and partial match, we also observe cases where the correctly predicted sentence extend into the neighboring sentence, whether at the beginning or at the end of the sentence. In other words, the predicted sentence contain parts of other sentence that it shouldn't. Out of 1378 sentences, we found 341 cases like this. It composed of 138 case where the sentence extend



at the beginning into prior sentence, 126 case where the sentence extend at the end into next sentence, and 70 where the sentence extend at the beginning and end. In average, it extend 9.2 token at the beginning and 16.7 token at the end of the sentence. More than 90% of the excessive tokens belong to background sentence, which means that background sentence are more likely to extend into neighboring sentence than other types of sentence.

We also observe cases where a sentence is not predicted, not even partially. We count 104 sentences that fall into this case.

Other than cases mentioned before, we also found cases where the sentence boundary is predicted correctly, but the category is predicted incorrectly. We count that 76 sentence fall into this category out of 1063 sentence that the boundary is predicted correctly (7.14%).

The result of each steps (sentence boundary detection and sentence type classification) in strategy B is shown in Table 5 and Table 6. In this result, we evaluate the two steps independently, which means that when we evaluate the sentence type classification, we assume that the boundary is correct. Additionally, when evaluating strategy B, we also compare it with baseline method proposed in Mahendra et al. [4] which applied using the same dataset we use in our experiment. Furthermore, to compare strategy B with strategy A, we also present the result of dependent pipeline in strategy B, which means the second step is obtained using the result of the first step. The result is shown in table 7.

**Table 5.** Performance of Sentence Boundary Detection on Strategy B

Tag	Precision		Recall		F1	
	Base	SL	Base	SL	Base	SL
BEGIN	0.85	0.88	0.67	0.78	0.75	0.83
NOT_BEGIN	0.96	0.98	0.99	0.99	0.98	0.98
Average / Total	0.95	0.97	0.96	0.97	0.95	0.97

**Table 6.** Performance of Sentence Type Classification on Strategy B

Sentence Type	Precision		Recall		F1	
	Base	SL	Base	SL	Base	SL
BACKGROUND	0.84	0.91	0.93	0.97	0.88	0.94
IGNORE	0.82	0.95	0.77	0.89	0.80	0.92
QUESTION	0.95	0.95	0.79	0.88	0.86	0.91
Average / Total	0.86	0.93	0.86	0.93	0.86	0.93

**Table 7.** Performance of Sentence Recognition on Strategy B

Tag	Precision	Recall	F1 Score
B-BACKGROUND	0.72	0.65	0.68
I-BACKGROUND	0.90	0.91	0.91
B-IGNORE	0.76	0.65	0.70
I-IGNORE	0.72	0.60	0.66
B-QUESTION	0.71	0.64	0.67
I-QUESTION	0.76	0.82	0.79
Average / Total	0.84	0.85	0.84

## 4.2 Ignored Word Recognition

We evaluate the proposed ignored word recognition with 10-fold cross-validation. The result is shown in Table 8. In our observation of the results, most of the incorrect prediction is caused by inabilities of our proposed model to recognize the difference between word as a subject and word as a greeting. To give more illustration, consider the following examples:

- **dokter** urologi menerangkan bahwa ginjal anak saya membengkak. (*urology doctor explain that my child’s kidney swell.*)
- apa obat yang cocok, **dokter**? (*what is the suitable medicine, doctor?*)

The word ”doctor” in the first example is a subject, but the word ”doctor” in the second example is a greeting word. We define greeting word, such as ”doctor” in the second example, as word that can be ignored.

**Table 8.** Performance of Ignored Word Recognition

Precision	Recall	F1-Score
0.94	0.86	0.90

## 5 Conclusion

We have proposed the use of sequence labeling approach to find question and contextual medical information from text, along with part of the text that can be ignored. In general, our module receive complex health question as an input and produce list of background, question, and ignored sentence as an output. We use linear-chain Conditional Random Field heavily as our model and analyze several strategy in our proposed method. Our evaluation show that the proposed sequence labeling method achieve 0.83 F1 score in detecting boundary of sentence and 0.93 F1 score when classifying type of sentence. Furthermore, when detecting word that could be ignored in sentence, the proposed method achieve 0.90 F1 score.

For future works, it is interesting to see the use of deep learning feature, such as word embedding, or other machine learning model, such as Structured SVM, Perceptron, Max Margin Markov, or Long Short-Term Memory (LSTM), in our proposed method. We also curious of how the model perform in a much larger size of dataset.

## Acknowledgement

The authors gratefully acknowledge the support of the PITTA UI Grant Contract No. 409/UN2.R3.1/HKP.05.00/2017. The first author was also partially funded by Bukalapak

## References

- [1] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung. Designing an indonesian part of speech tagset and manually tagged indonesian corpus. In *2014 International Conference on Asian Language Processing (IALP)*, pages 66–69, Oct 2014. doi: 10.1109/IALP.2014.6973519.
- [2] Kilian Evang, Valerio Basile, Grzegorz Chrupala, and Johan Bos. Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/D13-1146>.
- [3] Abid Nurul Hakim, Rahmad Mahendra, Mirna Adriani, and Adrianus Saga Ekakristi. Corpus development for indonesian consumer-health question answering system. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017.
- [4] Rahmad Mahendra, Abid Nurul Hakim, and Mirna Adriani. Towards question identification from online healthcare consultation forum post in bahasa. In *2017 International Conference on Asian Language Processing (IALP)*, 2017.
- [5] Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-fushman. Annotating question decomposition on complex medical questions. In *In Proceedings of LREC*, 2014.
- [6] Kirk Roberts, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. Decomposing consumer health questions. In *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing*, pages 29–37. U.S. National Library of Medicine, 2017.
- [7] Parikshit Sondhi, Manish Gupta, ChengXiang Zhai, and Julia Hockenmaier. Shallow information extraction from medical forum data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10*, pages 1158–1166, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL <http://dl.acm.org/citation.cfm?id=1944566.1944699>.