Journal of Biomedical Semantics

RESEARCH Open Access

Sentences, entities, and keyphrases extraction from consumer health forums using multi-task learning

Tsaqif Naufal[1], Rahmad Mahendra[1] and Alfan Farizki Wicaksono[1*]

Abstract

Purpose Online consumer health forums offer an alternative source of health-related information for internet users seeking specific details that may not be readily available through articles or other one-way communication channels. However, the effectiveness of these forums can be constrained by the limited number of healthcare professionals actively participating, which can impact response times to user inquiries. One potential solution to this issue is the integration of a semi-automatic system. A critical component of such a system is question processing, which often involves sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE) modules. We posit that the development of these three modules would enable the system to identify critical components of the question, thereby facilitating a deeper understanding of the question, and allowing for the re-formulation of more effective questions with extracted key information.

Methods This work contributes to two key aspects related to these three tasks. First, we expand and publicly release an Indonesian dataset for each task. Second, we establish a baseline for all three tasks within the Indonesian language domain by employing transformer-based models with nine distinct encoder variations. Our feature studies revealed an interdependence among these three tasks. Consequently, we propose several multi-task learning (MTL) models, both in pairwise and three-way configurations, incorporating parallel and hierarchical architectures.

Results Using F1-score at the chunk level, the inter-annotator agreements for SR, MER, and KE tasks were 88.61%, 64.83%, and 35.01% respectively. In single-task learning (STL) settings, the best performance for each task was achieved by different model, with IndoNLU$_{\text{LARGE}}$ obtained the highest average score. These results suggested that a larger model did not always perform better. We also found no indication of which ones between Indonesian and multilingual language models that generally performed better for our tasks. In pairwise MTL settings, we found that pairing tasks could outperform the STL baseline for all three tasks. Despite varying loss weights across our three-way MTL models, we did not identify a consistent pattern. While some configurations improved MER and KE performance, none surpassed the best pairwise MTL model for the SR task.

Conclusion We extended an Indonesian dataset for SR, MER, and KE tasks, resulted in 1, 173 labeled data points which splitted into 773 training instances, 200 validation instances, and 200 testing instances. We then used transformer-based models to set a baseline for all three tasks. Our MTL experiments suggested that additional information regarding the other two tasks could help the learning process for MER and KE tasks, while had only a small effect for SR task.

*Correspondence:
Alfan Farizki Wicaksono
alfan@cs.ui.ac.id

Keywords Consumer health question-answering system, Sentence recognition, Medical entity recognition, Keyphrase extraction, Multi-task learning

Introduction

The increasing accessibility of the internet leads to a growing number of people using it in their daily lives. A common online activity is seeking health-related information. While numerous articles exist, not all public needs are addressed. In such cases, individuals can consult doctors or healthcare professionals through online consumer health forums. However, questions asked by users are often unable to be answered immediately due to the high volume of incoming questions and the limited number of healthcare professionals. One solution to address this issue is by integrating a semi-automatic system [1–7]. However, the development of semi-automatic system in the healthcare field faces its own difficulties due to the limited available data and the complexity of the healthcare field itself [8]. Additionally, questions posed by the general public in consumer health forums are often written without adhering to standard rules. Figure 1 shows an example of question text from consumer health forum. There are several writing rules that were violated in the example, such as not starting sentences with capital letters, improper use of punctuation, and writing abbreviations in lowercase letters (i.e. "*tht*" instead of "*THT*").

A semi-automatic system varies widely depending on the desired functionality. Each system may consist of one or more modules to achieve its objective. In this research, we focus on developing three modules, i.e. sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE). As questions on consumer health forums are often written in narrative form, sentence recognition module can be used to split those questions into smaller segments. Each segment then can be determined whether it contains question or any other information that may be useful to help answer the question. The unimportant parts, such as preamble and greeting, can be omitted before proceeding to the next process. Meanwhile, medical entity recognition and keyphrase extraction modules are helpful to capture the focus of the questions. The output of both modules can also be used to formulate query, either to find similar previous questions or to retrieve relevant documents in the collection.

Despite existing research on these three modules in Indonesian, significant room for improvement remains. Regarding datasets, previous research has employed relatively small datasets, such as 192 instances for SR [9] and 309 instances for MER [10] and KE [11]. Moreover, annotation guidelines and inter-annotator agreement for these datasets were not disclosed. Methodologically, the models used in these studies were based on Conditional Random Fields (CRFs) and Long Short-Term Memories (LSTMs). Since these studies were conducted, there have been many developments in the field of Natural Language Processing (NLP), one of them is the development of Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. [12] and other transformer-based models. By adding an output layer on top of pretrained transformers and applying fine-tuning, the results obtained for various NLP tasks, such as question answering and language inference, have surpassed the state-of-the-art at that time. We believe that the use of

Selamat sore dokter. saya memang menderita polip dan sudah periksa di dokter tht, lalu saya dikasih obat terapi untuk 15 hari. habis obat kembali ke dokter polipnya mengecil. tetapi sekarang kok rasanya tersumbat lagi. yang ingin saya tanyakan apakah kemungkinan saya harus operasi atau hanya dikasih obat semprot bisa sembuh? terima kasih

English translation (for illustration):
Good afternoon, Doctor. I have been diagnosed with polyps and have already consulted an ENT specialist. I was prescribed a 15-day therapy medication. After finishing the medication and returning to the doctor, the polyps had shrunk. However, now I feel congested again. I would like to ask whether I might need surgery or if using a nasal spray alone could be enough for recovery. Thank you.

Fig. 1 Example of a question text from consumer health forum

transformer-based models can provide better results for these three modules.

While we are aware that generative large language models (LLMs), such as GPT [13] and LLaMA [14], have been rapidly advancing and have significantly transformed research in the field of NLP in recent years, we choose not to to explore that area yet. In addition to their higher computational costs, several studies have found that generative LLMs do not consistently outperform BERT-based models in certain settings [15–17]. Despite their advantages in zero-shot and few-shot scenarios, generative LLMs typically fail to surpass the performance of smaller models fine-tuned on a full dataset, particularly in tasks formulated as sequence labeling problem. Given that our dataset is quite sizable, we choose to prioritize the use of BERT-based models in this work.

The selection of the sentence recognition, medical entity recognition, and keyphrase extraction modules in this work is also based on the interdependence between these three modules. Medical entities present in a question often intersect with keyphrases from that question. A feature ablation study by Saputra et al. [11] also showed that the use of medical entities as a feature has a positive effect on model performance. In relation to the sentence recognition module, a sentence that contains medical entities or keyphrases have a higher probability of being a background or question compared to being an ignore-type. The interdependence between these three modules opens up opportunities for the application of multi-task learning (MTL), with the hope that information learned from one task can help the learning process of another task. Based on the preceding points, our research contributions are as follows:

1. **Dataset Expansion**: We have extended and published an Indonesian dataset for three tasks: sentence recognition, medical entity recognition, and keyphrase extraction;
2. **Baseline Establishment**: We have set a baseline for these tasks within the Indonesian domain;
3. **Multi-task Learning**: We have implemented multi-task learning to simultaneously address sentence recognition, medical entity recognition, and keyphrase extraction.

The remainder of this paper is structured as follows. "Backgrounds" section presents background information on the three tasks addressed in this study – sentence recognition, medical entity recognition, and medical keyphrase extraction – along with a review of existing methodologies. "Task definition" section provides the definition of the three tasks within the context of this research. "Data annotation" section details our efforts to expand and enhance existing datasets for these tasks. "Task modeling" section outlines our proposed approach, including our proposed multi-task learning, to tackle the three tasks. "Results and analysis" section presents our experimental results and a comprehensive analysis. "Conclusion" section concludes the paper, summarizing key findings and contributions.

Backgrounds

In this work, we address three tasks: sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE). Other than their standalone functionalities, these tasks are also essential for the question processing module in a semi-automatic consumer health system. "Sentence recognition" section discusses sentence recognition, including related work on English and Indonesian consumer-health documents. "Medical entity recognition" section explores methods for extracting medical entities from health-related documents. This is essential for identifying the specific topic or condition of a user's inquiry. In conjunction with keyphrase extraction ("Keyphrase extraction" section), medical entity recognition identifies details such as symptoms, medications, or procedures. These details facilitate integration with medical knowledge bases, enabling the system to provide more contextual and authoritative responses.

Sentence recognition

A question text on a consumer health forum is typically a lengthy text composed of several sentences. Roberts et al. [18] described that kind of question as a complex question, which consists of more than one sentence, contains background information, and generally includes more than one specific question. A complex question can be decomposed into more than one specific question, where each question can be answered independently. In their research on question decomposition, Roberts et al. [18] proposed six components, including sentence recognition (SR). SR nvolves splitting sentences and classifying them by type. They defined three sentence types: background, question, and ignore. The proposed model, a Support Vector Machine (SVM), utilized unigrams, bigrams, parse tree tags, and part-of-speech (POS) features.

On the other hand, Mahendra et al. [19] explored sentence recognition in Indonesian consumer health forums, treating sentence splitting and classification as separate modules. In sentence splitter component, a question text is first broke into text segments, then heuristic rules based on the appearance of punctuations are applied. For sentence classification, they used SVM with four features groups, i.e. n-grams, position and length of sentence, question-specific attributes, and dictionary of symptoms

and diseases. Ekakristi et al. [9] also had SR as one of their components and proposed two strategies to solve it. The first one is by predicting the boundary and type of each token in one sequence labeling process, while the second one is by determining the boundary of sentences at the beginning before assigning the type of each sentence afterwards. Using hand-crafted features and CRFs as models, they found that the first strategy achieved a better performance.

Medical entity recognition

Medical entity recognition (MER), also known as clinical named entity recognition, is a branch of named entity recognition (NER) that focuses on the health domain. In a semi-automatic system, extracting valuable information from the data is one of the most important initial steps. The outcome from MER module, either applied to question text or document collection, then can be used to support subsequent processes, such as document indexing and entity linking.

Abacha and Zweigenbaum [20] compared three methods based on domain knowledge and machine learning techniques: (i) semantic method using MetaMap, (ii) chunker-based noun phrase extraction followed by classification using SVM, and (iii) hybrid approach using CRFs with semantic rule as additional features. Their evaluations showed that the hybrid approach obtained the best performance among the three. Wu et al. [21] examined the utilization of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to extract concepts from clinical texts and found that RNN achieved a better performance. Xu et al. [22] developed a model based on bidirectional LSTM (BiLSTM) and CRF, where two BiLSTM layers are used to obtain character and word embedding. Cho et al. [23] also proposed BiLSTM-CRF model, with the addition of CNN to obtain character embedding and an attention layer before entering the CRF layer.

In recent years, several works utlized BERT to extract medical entities, such as Yu et al. [24] which used BioBERT [25], Peng et al. [26] which pre-trained BERT on PubMed abstracts and MIMIC-III clinical notes, and Dai et al. [27] which combined BERT with BiLSTM-CRF model. Some studies also utilized generative LLMs. GPT-NER [15] transformed the task of finding location entities in the text to the task of generating text sequence with special tokens to mark entities. PromptNER [28] included a modular definition of entity types as one of its components, in addition to the few-shot examples from the target domain. Hu et al. [16] compared the performance of ChatGPT, GPT- 3 [13], and BioClinicalBERT [29] on clinical NER tasks. Their experiments revealed that ChatGPT achieved reasonable performance on zero-shot setting, but it was still not as good as fine-tuned BioClinicalBERT.

For MER in Indonesian language domain, Suwarningsih et al. [30] used SVM with word-level, list, and corpus features to extract medical entities from Indonesian medical articles. Sadikin et al. [31] focused on drug entities and proposed three data representation techniques based on the characteristics of word distribution and word similarities as a result of word embedding training. For modeling purpose, they used standard neural networks model, Deep Belief Network (DBN), Stacked Autoencoder (SAE), and LSTM. They found that LSTM rendered the best result for their case. In other works, both Herwando et al. [32] and Rohman [10] defined four medical entity types, i.e. disease, symptom, drug, and treatment. Using hand-crafted features, Herwando et al. [32] used CRFs as the model while Rohman [10] used LSTM-based models.

Keyphrase extraction

Keyphrase is a concise expression that can represent important information within a document. The main objective of keyphrase extraction (KE) task is to automatically extract a set of representative phrases that concisely summarize the content of a document [33]. Keyphrase extraction has been applied to various type of texts, ranging from long documents (e.g. web pages [34] and scientific articles [35, 36]) to user-generated contents (e.g. tweets [37], e-mail [38], and chats [39]). Extracting keyphrases from user-generated contents, including questions on consumer health forum, has its own difficulty due to unstructured format and typographical mistakes. In long text such as a request on consumer health forum, the ability to extract keyphrases makes it easier to capture the essence of the questions. The extracted keyphrases then can be used to formulate a query to retrieve similar questions or information needed to generate an answer from a collection of medical documents.

In general, KE methods can be categorized into unsupervised and supervised approach. While unsupervised methods are domain independent and do not need labeled training data, supervised methods have more powerful modeling capabilities and typically perform better [40]. Unsupervised methods include YAKE [41], TextRank [42], and RAKE [43]. Meanwhile, supervised methods include traditional (e.g. KEA [44]) and deep learning [45–47] methods. Several works in recent years formulated keyphrase extraction as generative process. KeyBART [48] continued the training of BART [49] by performing token masking, keyphrase masking, and keyphrase replacement on the input text and pre-trained the model to predict the original keyphrases in CatSeq format. Zhu et al. [50] employed a non-autoregressive

decoder to generate all possible keyphrases, both present and absent, in parallel. Other studies proposed integrated models that combined extractive and generative approaches. Chen et al. [51] proposed a multi-task learning framework with a neural-based merging module to combine and re-rank the predicted keyphrases from the enhanced generative model, the extractive model, and the retrieved keyphrases. UniKeyphrase [52] incorporated stacked relation layer to capture relation between present keyphrase extraction (PKE) and absent keyphrase generation (AKG), also bag-of-words constraint to feed global information about present and absent keyphrases to the model.

For keyphrase extraction in medical domain, Cao et al. [53] used logistic regression and CRFs as models with n-grams, POS tag, stemming, word length, and word position as features. Their evaluations showed that both models outperformed unsupervised baselines with CRFs achieved a higher F1-score. Sarkar [54] applied a hybrid method that combines statistical and knowledge base methods to assign weights to candidate keyphrases. To identify candidate keyphrases, they used punctuations and stopwords as splitting point. Recent works on this topic utilize BERT-based models to extract keyphrases from medical documents [55, 56]. For Indonesian language domain, Saputra et al. [11] used hand-crafted features and LSTM-based models to extract keyphrase from questions in consumer health forums.

Task definition

Sentence Recognition (SR). Following [9, 18, 19], we define three types of sentences in this work, i.e. `Background`, `Question`, `Ignore`. A `Background` sentence is defined as a sentence that contains useful contextual information that does not include a question. Meanwhile, a `Question` sentence is defined as a sentence that contains one or more question clauses. Lastly, an `Ignore` sentence is defined as a sentence that can be disregarded because it does not contain any relevant information and/or question, including preamble, greeting, general intention, general request, gratitude, and courtesy. Given an input in the form of a question from consumer health forums, the expected output from the SR module is the sentences identified in that question along with their respective types. Figure 2 shows an example of question text from consumer health forum along with the type of each sentence within it.

Medical Entity Recognition (MER). Following [10, 32], we define four medical entity types in this work, i.e. `Disease`, `Symptom`, `Drug`, and `Treatment`. The `Disease` entity refers to the name of an abnormal condition that arises in the human body, both physically and mentally. The `Symptom` entity refers to indication or circumstance experienced by someone who is affected by a disease. The `Drug` entity refers to the name of medicine which has the function of preventing, reducing, or curing the disease. The `Treatment` entity refers to a method or procedure to remediate a health problem. Given an input in the form of a questions from consumer health forums, the expected output from the MER module is a list of medical entities present in that question. Figure 3 shows an example of question text from consumer health forum along with medical entities within it.

Keyphrase Extraction (KE). A keyphrase is defined as one or more sequential words that provide important information and describe the essence of a document. One document can have multiple keyphrases. In this work, we limit the scope to extractive keyphrases, i.e. keyphrases that written explicitly in the text. Given an input in the form of a questions from consumer health forums, the expected output from the KE module is a list of keyphrases representing the essence of that question. Figure 4 shows an example of question text from consumer health forum along with keyphrases within it.

Dok, apakah dispepsia bisa menyebabkan nyeri punggung dan nyeri dada juga? Karena saya didiagnosa menderita dispepsia, namun punggung dan dada saya sering nyeri juga. Apakah saya perlu periksa ke dokter lain? Mohon masukkannya. terima kasih.

English translation (for illustration):
Doctor, can dyspepsia cause back pain and chest pain as well? I have been diagnosed with dyspepsia, but I also often experience pain in my back and chest. Should I see another doctor? Please give me an advice. Thank you.

Fig. 2 Example of a consumer health question as input for sentence recognition task. indicates `Background` sentence, indicates `Question` sentence, and indicates `Ignore` sentence

Pagi dok. Saya terkena infeksi kelenjar getah bening di leher dan dokter memberi obat pada saya. Kalau obat sudah habis leher saya kaku dan kepala terasa nyeri. Dokter saya memberikan obat yang namanya cataflam, sanexon, dan bicrolid. Apakah saya harus berhenti minum obat yang dikasih dokter dan ganti dengan kompres air hangat saja? Terimakasih

English translation (for illustration):
Good morning, Doctor. I have a lymph node infection in my neck, and the doctor gave me medication. When the medication runs out, my neck becomes stiff and I feel pain in my head. My doctor prescribed me cataflam, sanexon, and bicrolid. Should I stop taking the medication prescribed by the doctor and switch to using warm compress instead? Thank you.

Fig. 3 Example of a consumer health question as input for medical entity recognition task. indicates `Disease` entity, indicates `Symptom` entity, indicates `Drug` entity, and indicates `Treatment` entity

Dok, waktu saya pergi urut, tukang urutnya bilang ada pembengkakan perut dari tengah di atas pusat terus ke bawah sebelah kiri. Dan saya juga sering ada benjolan seperti bisul kecil di kelopak mata. Dan kedua penyakit ini saya rasakan sekitar 2 sampai 3 bulan belakangan ini. Walau tidak terasa sakit tetapi belakangan badan saya sering terasa lemas. Apakah kedua penyakit ini ada hubunganya? Apa nama penyakit saya ini kanker? Dan seberapa parahkah penyakit ini? terima kasih.

English translation (for illustration):
Doctor, when I went for a massage, the masseur said there was swelling in my abdomen from the middle above the navel down to the left side. I also frequently have small bumps like boils on my eyelids. I have been experiencing these two conditions for about the past 2 to 3 months. Although they are not painful, my body often feels weak lately. Are these two conditions related? Is this cancer? How serious is this condition? Thank you.

Fig. 4 Example of a consumer health question as input for keyphrase extraction task. indicates `Keyphrase`

Data annotation

The datasets used for sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE) are from Ekakristi et al. [9], Rohman [10], and Saputra et al. [11], respectively. Specifically, we identified 173 overlapping samples from the 192 originally used for SR and the 309 used for both MER and KE. These 173 samples shared the same text but required label adjustments to ensure consistency with our newly developed annotation guidelines.

Recognizing the limitations of the existing datasets, we performed supplemental annotation. Specifically, we extracted one thousand data points from the Hakim et al. [57] corpus that had not been previously utilized. The entire annotation process was conducted using Label Studio [58] as our annotation tool. For each of the three tasks, we engaged two annotators with prior experience in data annotation. Due to resource constraints (budget and time), our annotation process was divided into two phases: **paired annotation**, where each data instance was annotated by two annotators, and **single annotation**, where an instance was labeled by only one annotator.

During the **paired annotation** phase, each annotator was assigned 400 data points. Adhering to the methodology outlined in Lehman et al. [59], our annotation process comprised pilot and final rounds. Initially, we provided the annotators with annotation guidelines and instructed them to annotate the data accordingly. Subsequently, we calculated the inter-annotator agreement. We proceeded to solicit feedback from the annotators and analyze the annotation discrepancies observed during the pilot round. We observed that, in addition

to insufficient instruction clarity, several annotation discrepancies arose due to annotator oversights, such as unlabeled tokens in the SR task and omitted entities in the MER task. We refined the annotation guidelines and provided feedback to the annotators regarding common mistakes identified during the pilot round. The final annotation guidelines for all tasks are available in Appendix A. We instructed the annotators to revisit and revise their annotations in accordance with the improved guidelines. Finally, we recalculated the inter-annotator agreement upon completion of the final annotation round.

Various methods exist to evaluate inter-annotator agreement, with Cohen's Kappa [60] being the most frequently reported measure in medical literature [61]. Cohen's Kappa agreement is calculated using the following equation:

$$K = \frac{P_a - P_e}{1 - P_e}, \tag{1}$$

where P_a denotes the probability of actual agreement and P_e denotes the probability of expected agreement. However, several studies [62, 63] have noted that Cohen's Kappa is not optimal for sequence labeling tasks. Therefore, following Deleger et al. [64] and Brandsen et al. [65], we opted to use the F1-score at the chunk level, in conjunction with Cohen's Kappa, to assess inter-annotator agreement.

Table 1 displays the inter-annotator agreement scores for all three tasks, comparing the initial and final rounds. We observe an increase in the metrics, signifying an enhancement in the reliability of the annotated data. While a degree of disagreement persists for KE, which proved to be the most difficult, the overall reliability improved."

In the **single annotation** phase, each annotator worked on 300 unique data instances. Although the annotations in this phase were not fully cross-checked, we believe they are reliable due to the annotators' proficiency gained during the pilot and final rounds of the paired annotation phase.

Table 1 Inter-annotator agreement for sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE) dataset

Phase	Agreement metrics	SR(%)	MER(%)	KE(%)
Initial annotation	F1-score	72.84	49.41	28.89
	Cohen's Kappa	84.85	63.98	31.51
Final annotation	F1-score	88.61	64.83	35.01
	Cohen's Kappa	94.63	76.51	44.85

As we aim to preserve all annotated data as a gold standard, we addressed the instances of label disagreement that remained after the final round of paired annotation. Therefore, we devised rule-based adjudication methods to determine the final labels, which are detailed as follows.

1. In cases where one annotator labels a chunk as a single sentence and the other annotator identifies two different sentences of the same type within that chunk:
   - For background sentences, label them as two different sentences;
   - For question sentences, label them as two different sentences if both chunks can stand alone;
   - For ignore sentences, label them as a single sentence considering that ignore sentences will not be used in subsequent stages;
2. In cases where one annotator labels a chunk as a single sentence and the other annotator identifies two different sentences of different types within that chunk, label them as two sentences with different types as long as it does not seem unusual;
3. In cases where both annotators label the same chunk with different sentence types, the final label will be determined by the author as the third annotator, prioritizing background and question sentence types;
4. In cases where annotators identify two chunks with different boundaries and one chunk is labeled as an ignore sentence, choose the annotation with the better quality of background or question sentence;
5. For other annotation disagreement, the author as the third annotator will determine the final label.

For MER and KE tasks, to address cases where annotators disagree on the boundaries of entities or keyphrases, the intersection of both annotation results will be used as the gold standard. For other annotation disagreements (e.g. disagreements on entity types or when only one annotator labels a span as an entity or keyphrase), the first author, acted as the third annotator, determine the final label.

Furthermore, to ensure the quality of the dataset, we conducted a sample check of the annotation results from the single annotation phase. Any annotations that demonstrably violated the annotation guidelines were removed from the gold standard.

The time required to finish the annotation process varies for each annotator, ranging from 18 to 25 hours that spread over three months. In the end, there are 1, 173 labeled data after the annotation process was

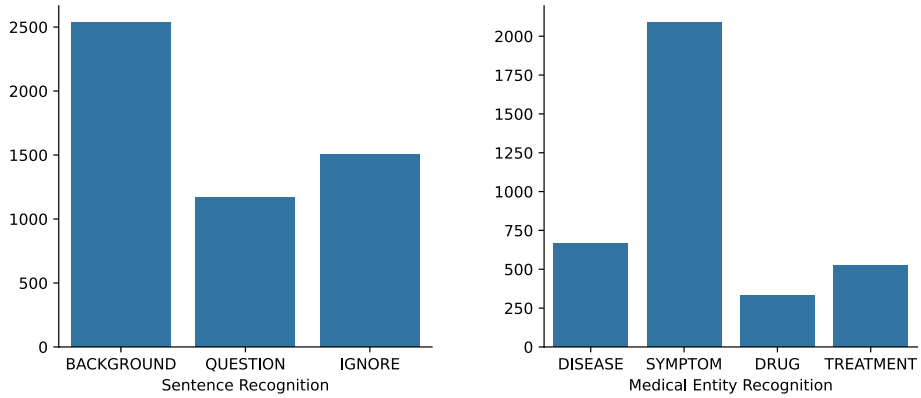

Fig. 5 Distribution of sentence types for sentence recognition (left) and entity types for medical entity recognition (right) in the training data

Table 2 Ratio of each sentence type at the token level based on MER and KE tags in the training data

Entity type	Keyphrase?	Sentence type (%)		
		Background	Question	Ignore
Disease	✓	58.41	40.27	1.33
	×	75.71	20.00	4.29
Symptom	✓	93.59	6.41	0.00
	×	97.98	2.02	0.00
Drug	✓	74.62	25.38	0.00
	×	89.83	10.17	0.00
Treatment	✓	63.14	36.83	0.00
	×	90.11	9.89	0.00

completed. We then divided them into three sets: 773 instances for training, 200 instances for validation, and 200 instances for testing. The training data consists of a combination of 173 instances from the adjusted initial dataset and 600 instances from the third annotation phase. Meanwhile, 400 instances from the first and second annotation phases will be randomly split for validation and testing. Overall, there are 2, 987 keyphrases in the training data, while the distribution of sentence types and entity types can be seen in Fig. 5.

Table 2 presents the ratio of each sentence type (i.e. SR tags) at the token level based on entity type (i.e. MER tags) and whether it is a part of keyphrase (i.e. KE tags) in the training data. There are no tokens labeled as part of symptom, drug, or treatment entity that are also part of ignore sentence. For disease entity, tokens that labeled as part of ignore sentence are typically found in the following kind of sentence: "*Dok, saya ingin bertanya tentang <DISEASE>*" ("Doctor, I would like to ask about <DISEASE>"). This observation suggested that a sentence that contains medical entities or keyphrases tends to be a background or question rather than ignore-type. Furthermore, we also investigated whether the presence of medical entities or keyphrases in a sentence is correlated to the type of sentence. Table 3 presents the number of sentences containing each medical entity types and keyphrases for each sentence type. Using chi-square test with significance level of 0.05, we can conclude that there is a relationship between the presence of medical entities or keyphrases with sentence type (p-value < 0.05).

Task modeling

We formulate sentence recognition, medical entity recognition, and keyphrase extraction tasks as sequence labeling problem, which aims to label each token in a sequence. The input and output of sequence labeling can be expressed as

Table 3 The number of sentences containing medical entities and keyphrases for each sentence type. 'None' columns shows the number of sentences that do not contain any medical entities or keyphrases

Sentence type	MER					KE	
	Disease	Symptom	Drug	Treatment	None	Keyphrase	None
Background	330	1, 295	166	348	718	1, 565	974
Question	252	114	54	105	705	546	623
Ignore	8	0	0	0	1, 500	9	1, 499

$$x = (x_1, x_2, ..., x_N),$$

$$y = (y_1, y_2, ..., y_N),$$

where x and y are sequences. The notation y_i represents the label corresponding to x_i, where $y_i \in C$ and C denotes a finite set of labels.

For all three tasks, x represents a sequence of tokens in a question posted on a health QA forum, where x_i denotes the token at position i. The difference among these three tasks lies in the finite label sets used to label each token in x. As we use BIO (Begin, Inside, Outside) tagging format, a token is labeled as `B-` if it is a beginning of a chunk, `I-` if it is a part of a chunk but not located at the beginning, and `O` if it is not a part of any chunks. For the sentence recognition task, $y_i \in$ {`B-BACK.`, `I-BACK.`, `B-QUE.`, `I-QUE.`, `B-IGN.`, `I-IGN.`} where `BACK.`, `QUE.`, and `IGN.` respectively indicate the types of background, question, and ignore sentences. For the medical entity recognition task, $y_i \in$ {`B-DIS.`, `I-DIS.`, `B-DRUG`, `I-DRUG`, `B-SYMP.`, `I-SYMP.`, `B-TREAT.`, `I-TREAT.`, `O`} where `DIS.`, `DRUG`, `SYMP.`, and `TREAT.` respectively indicate disease, drug, symptom, and treatment entities. For the keyphrase extraction task, $y_i \in$ {`B-KEY`, `I-KEY`, `O`}. Table 4 shows examples of x and y for the three tasks mentioned.

Various types of models can be used to tackle a sequence labeling problem. In this work, we propose the utilization of transformer-based models for both single-task and multi-task learning approaches. The results obtained then were compared to CRFs as the baseline.

Conditional random fields

Conditional Random Fields (CRFs) [66] can be viewed as a log-linear model that calculates the probability of a tag sequence y from all of possible tag sequences when given a sequence x. The specific and common form of CRFs used for sequence labeling is the linear-chain CRFs. Given the input sequence $x = (x_1, x_2, ..., x_N)$ and the target sequence $y = \{y_1, y_2, ..., y_N\}$, the linear-chain CRF models the conditional probability as

$$P(y|x) = \frac{\exp\left(\sum_{i=1}^{N} U(x_i, y_i) + \sum_{i=0}^{N} T(y_i, y_{i+1})\right)}{Z(x)}, \tag{2}$$

where $U(x_i, y_i)$ is the emissions score for x_i with label y_i; $T(y_i, y_{i+1})$ is the transition score between label y_i and y_{i+1}; and $Z(x)$ is a normalization factor. The value of x can be a handcrafted features or the output from previous layers if we put CRFs on top of neural networks. The loss then can be obtained by calculating the negative log-likelihood as follows:

$$\mathcal{L}_{CRF} = -\log P(y|x)\,. \tag{3}$$

Building upon previous studies in the Indonesian language domain that also utilized data from online consumer health forums (i.e. Ekakristi et al. [9] for SR, Rohman [10] for MER, and Saputra et al. [11] for KE), this work employs CRFs as the baseline model. Since CRFs model is computationally less expensive than our proposed transformer-based models, utilizing it to address the three tasks might be more favorable if its performance is not significantly inferior. We train the CRFs model by incorporating handcrafted features that were previously used in those three studies with some adjustments due to the differences in data characteristics. List of handcrafted features used for each task can be seen in Table 5.

Transformer-based models

When using transformer-based models, there is a chance that a token is divided into subtokens during the tokenization process. We also need to add special token `[CLS]` at the beginning and `[SEP]` at the end of each input. In order to handle this issue, the input x is transformed into $T = \{t_{[\text{CLS}]}, t_1, t_2, ..., t_N, t_{[\text{SEP}]}\}$ before being fed to the models, where $t_i = [t_{i,1}, t_{i,2}, ..., t_{i,m}]$ with m is the number of subtokens of token t_i. The output of encoder then can be expressed as $\mathbf{W} = \{\mathbf{w}_{[\text{CLS}]}, \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N, \mathbf{w}_{[\text{SEP}]}\}$, where $\mathbf{w}_i$ is last hidden state of t_i. In cases where t_i is splitted into subtokens, the value of $\mathbf{w}_i$ is obtained from the last hidden state of the first subtoken, i.e. $\mathbf{w}_{i,1}$. The

Table 4 Example of input and output for sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE) tasks using sequence labeling approach

x	siang	dok	,	apa	penyebab	sinusitis
y_{SR}	B-IGN.	I-IGN.	I-IGN.	B-QUE.	I-QUE.	I-QUE.
y_{MER}	O	O	O	O	O	B-DIS.
y_{KE}	O	O	O	O	B-KEY	I-KEY

Table 5 List of handcrafted features used for each task

Feature	Explanation	SR	MER	KE
token	current token	✓	✓	✓
token_before	previous token	✓	✓	×
token_after	next token	✓	✓	×
is_digit	whether current token is digit	✓	×	×
is_begin	whether current token is the beginning of input text	✓	×	×
is_end	whether current token is the end of input text	✓	×	×
token_before_is_closure	whether previous token is one of the following characters:. (period),! (exclamation mark),? (question mark), or, (comma)	✓	×	×
token_length	length of current token	×	×	✓
absolute_position	index of current token, starting from 0	×	×	✓
relative_position	index of current token divided by total number of tokens in the input	✓	×	×
site_code	code of the website from which the input text was obtained	✓	×	×
pos_tag	part-of-speech tag for current token	✓	×	✓
is_np	whether current token is part of a noun phrase	×	✓	×
is_vp	whether current token is part of a verb phrase	×	✓	×
is_stopword	whether current token is a stopword	×	✓	×
is_abbreviation	whether current token is an abbreviation	×	×	✓
abbreviation_inverse	full form of current token if it is an abbreviation	✓	×	×
is_disease	whether current token is in disease dictionary	×	✓	×
is_symptom	whether current token is in symptom dictionary	×	✓	×
is_treatment	whether current token is in treatment dictionary	×	✓	×
in_medical_dict	whether current token is in one of the disease, drug, symptom, or treatment dictionaries	×	×	✓
is_medical_entity	whether current token is part of a medical entity	×	×	✓
max_lcs_disease	maximum ratio of longest common substring between current token and entries in disease dictionary	✓	×	×
max_lcs_symptom	maximum ratio of longest common substring between current token and entries in symptom dictionary	✓	×	×
max_lcs_treatment	maximum ratio of longest common substring between current token and entries in treatment dictionary	✓	×	×
stickiness	stickiness value between cureent token with its preceding and succeeding tokens, computed using point-wise mutual information (PMI)	×	×	✓

last hidden state of special tokens, i.e. $\mathbf{w}_{\texttt{[CLS]}}$ and $\mathbf{w}_{\texttt{[SEP]}}$, is omitted before entering the next layer, results in $\mathbf{W}' = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N\}$. This entire process was utilized for all types of transformer-based models in this work.

Single-task Learning. We employ various kinds of pretrained transformers as encoder for single-task learning (STL) approach. For the Indonesian language models, we use IndoDistilBERT[1], $\text{IndoBERT}_{\text{BASE}}$ by IndoLEM [67] (denoted by $\text{IndoLEM}_{\text{BASE}}$), both the base and large variations of IndoBERT by IndoNLU [68] (denoted by $\text{IndoNLU}_{\text{BASE}}$ and $\text{IndoNLU}_{\text{LARGE}}$ respectively), IndoBERTweet [69], and $\text{IndoNLU}_{\text{BASE}}$ that has been fine-tuned on MER task[2] (denoted by $\text{IndoNLU}_{\text{BASE}}$ FT). For multilingual language models, we use XLM-MLM [70], and both the base and large variations of XLM-R [71] (denoted by $\text{XLM-R}_{\text{BASE}}$ and $\text{XLM-R}_{\text{LARGE}}$ respectively). The processed output from encoder, i.e. $\mathbf{W}'$, is directly fed to the softmax layer to obtain a probability distribution (i.e. a simplex) over all possible tags.

Pairwise Multi-task Learning. We propose two kind of model architecture for multi-task learning (MTL) approach, i.e. parallel and hierarchical. The difference between the two lies in whether the output of one task is directly used for another task. The parallel architecture only shares encoder between tasks, while the output of last layer for one task in hierarchical architecture is also be fed to the output layer of another task. For parallel architecture, the probability distribution $\mathbf{p}_x$ for task x is obtained by:

$$\mathbf{p}_x = \texttt{softmax}(\texttt{FFNN}_x(\mathbf{W}')), \tag{4}$$

where $\texttt{FFNN}_x(.)$ is a feedforward neural network exclusively used for task x; and $\texttt{softmax}(z_i) = \exp(z_i)/\sum_j \exp(z_j)$. For hierarchical

[1] https://huggingface.co/cahya/distilbert-base-indonesia

[2] https://huggingface.co/stevenwh/indobert-base-p2-finetuned-mer-80k/

architectures, consider two tasks: A and B. The probability distribution for task A is obtained by

$$\mathbf{p}_A = \texttt{softmax}(\texttt{FFNN}_A(\mathbf{W}')) . \quad (5)$$

We then calculate the embedding of task A's tag using:

$$\mathbf{e} = \mathbf{L} \times \mathbf{p}_A , \quad (6)$$

where $\mathbf{L}$ is a trainable label weight matrix corresponds to task A. Lastly, the probability distribution for task B can be obtained as follows:

$$\mathbf{p}_B = \texttt{softmax}(\texttt{FFNN}_B([\mathbf{W}'; \mathbf{e}])) , \quad (7)$$

where $[a; b]$ is a concatenation between two vector: a and b. The loss for pairwise MTL ($\mathcal{L}_p$) can be calculated using a weighted linear combination of task A's loss ($\mathcal{L}_A$) and task B's loss ($\mathcal{L}_B$) with a balancing hyper-parameter $0 < \alpha < 1$:

$$\mathcal{L}_p = \alpha \, \mathcal{L}_A + (1 - \alpha) \, \mathcal{L}_B. \quad (8)$$

Three-way Multi-task Learning. Similar to pairwise MTL, we propose both parallel and hierarchical architectures for three-way MTL. For parallel architecture, the probability distribution $\mathbf{p}_x$ for task x in three-way MTL can also be obtained using Equation 4. For hierarchical architecture, we first obtain the probability distribution for MER task using following equation, where $\texttt{FFNN}_M$ is a feedforward neural network exclusively used for MER task:

$$\mathbf{p}_M = \texttt{softmax}(\texttt{FFNN}_M(\mathbf{W}')). \quad (9)$$

We then calculate the embedding of MER tag for KE task as follows, where $\mathbf{L}_{MK}$ is a trainable label weight matrix corresponds to MER task used exclusively for KE task:

$$\mathbf{e}_{MK} = \mathbf{L}_{MK} \times \mathbf{p}_M . \quad (10)$$

The probability distribution for keyphrase extraction task then can be obtained using the following equation, where $\texttt{FFNN}_K$ is a feedforward neural network exclusively used for KE task:

$$\mathbf{p}_K = \texttt{softmax}(\texttt{FFNN}_K([\mathbf{W}'; \mathbf{e}_{MK}])). \quad (11)$$

For SR task, we utilize two embeddings corresponding to MER and KE tasks. The embedding of MER tag ($\mathbf{e}_{MS}$) and KE tag ($\mathbf{e}_{KS}$) is computed as follows, where $\mathbf{L}_{MS}$ is a trainable label weight matrix corresponds to MER task used exclusively for SR task and $\mathbf{L}_{KS}$ is a trainable label weight matrix corresponds to KE task:

$$\mathbf{e}_{MS} = \mathbf{L}_{MS} \times \mathbf{p}_M , \qquad \mathbf{e}_{KS} = \mathbf{L}_{KS} \times \mathbf{p}_K . \quad (12)$$

Lastly, the probability distribution for sentence recognition task can be obtained as follows, where $\texttt{FFNN}_S$ is a feedforward neural network exclusively used for SR task:

$$\mathbf{p}_S = \texttt{softmax}(\texttt{FFNN}_S([\mathbf{W}'; \mathbf{e}_{MS}; \mathbf{e}_{KS}])). \quad (13)$$

The loss for three-way MTL ($\mathcal{L}_t$) can be calculated by following equation, where $\mathcal{L}_S$ is the loss for sentence recognition, $\mathcal{L}_M$ is the loss for medical entity recognition, $\mathcal{L}_K$ is the loss for keyphrase extraction, $0 < \alpha, \ \beta < 1$ and $\alpha + \beta < 1$:

$$\mathcal{L}_t = \alpha \, \mathcal{L}_S + \beta \, \mathcal{L}_M + (1 - \alpha - \beta) \, \mathcal{L}_K. \quad (14)$$

Results and analysis

Experimental Setting. We ran all of our experiments on a single NVIDIA DGX A100 GPU. We used Adam as optimizer with an initial learning rate of 5×10^{-5}, a batch size of 16, and trained the models for 30 epochs. The model at the end of iteration where the highest score on validation set was obtained will be used to evaluate model performance on the test set. For each model, we ran the experiments for five times and reported the average scores. Since the performance of deep neural networks highly dependent on hyperparameters, random seeds, and other stochastic factors, comparing the average scores of two models across several runs might not be enough to decide which model is better. Therefore, we also conducted statistical significance test using Almost Stochastic Order (ASO) [72, 73] with a significance level of 0.05.

Evaluation Metrics. We used F1-score to evaluate model performance in this work. For sentence recognition, the evaluation process is applied at token level. We report micro-averaged F1-score for sentence recognition due to class imbalance between `B-` and `I-` tags, as well as the higher significance of sentence type compared to sentence boundaries. For medical entity recognition and keyphrase extraction, the evaluation process follows the method described in the CoNLL evaluation script [74], where a predicted chunk is correct only if it is an exact match of the corresponding chunk in the gold standard. For example, if "*kanker darah*" ("blood cancer") in "*pasien itu menderita kanker darah*" ("the patient was suffering from blood cancer") sentence is labeled as disease, both "*kanker*" ("cancer") and "*menderita kanker darah*" ("suffering from blood cancer") as predicted disease entity is not counted as true positive since both of them are only partially match with the gold standard. For these two tasks, we report the macro-averaged F1-score.

Table 6 presents the results for all three tasks using various single-task learning (STL) models. All transformer-based models, regardless of the encoder, achieved higher scores than CRFs for all three tasks. However, there is no single model that consistently performed better than the others, with the highest F1-score achieved by XLM-R_{BASE} for sentence recognition, IndoNLU$_{\text{LARGE}}$

Table 6 Evaluation results of sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE) using single-task learning approach. † indicates a stochastically dominant performance over other models

Model/Encoder	SR	MER	KE
CRFs	64.61	36.75	19.87
IndoDistilBERT[1]	92.35	54.43	42.64
$\text{IndoLEM}_{\text{BASE}}$ [67]	93.26	56.93	† **47.48**
$\text{IndoNLU}_{\text{BASE}}$ [68]	92.08	54.58	43.46
$\text{IndoNLU}_{\text{BASE}}$ FT[2]	92.14	54.07	43.44
IndoBERTweet [69]	92.77	53.25	42.63
$\text{IndoNLU}_{\text{LARGE}}$ [68]	92.81	**59.59**	46.91
XLM-MLM [70]	90.21	54.81	38.40
$\text{XLM-R}_{\text{BASE}}$ [71]	† **93.70**	45.78	46.55
$\text{XLM-R}_{\text{LARGE}}$ [71]	93.37	59.32	43.03

[1] https://huggingface.co/cahya/distilbert-base-indonesia

[2] https://huggingface.co/stevenwh/indobert-base-p2-finetuned-mer-80k/

for medical entity recognition, and $\text{IndoLEM}_{\text{BASE}}$ for keyphrase extraction. Except when comparing the performance of $\text{IndoNLU}_{\text{LARGE}}$ with $\text{XLM-R}_{\text{LARGE}}$ for MER task, the score distribution of those three models for each respective task are stochastically dominant over other models (ASO, $\epsilon_{\text{min}} < 0.5$). These results did not indicate which one is generally better between Indonesian and multilingual language model for our tasks. We also could conclude that a larger model did not always perform better than the smaller ones in our settings.

For our multi-task learning (MTL) models, we used encoder that achieved the highest average score between all three tasks in the STL approach, i.e. $\text{IndoNLU}_{\text{LARGE}}$. For each pairwise combination of tasks, we tested three different α values for both parallel and hierarchical architecture. Table 7 presents the results of our pairwise MTL experiments, with the first task mentioned was treated as task A. When using parallel architecture, the performance of task A consistently became better while the performance of task B became worse as the α value got higher. The similar pattern was not clearly observed when using hierarchical architecture. We attributed this finding to the relationship between two tasks that was more directly connected in hierarchical architecture compared to the parallel one. For SR, the Parallel MER - SR model with α value of 0.3 is not stochastically dominant over the STL baseline even though it achieved a higher score. While we found that some configurations managed to achieve better scores for MER than the STL baseline, none of them is stochastically dominant. For KE, the Hierarchical MER - KE model with α value of 0.3 is stochastically dominant over the STL baseline and other pairwise MTL models, except the Parallel MER - KE model with α value of 0.3. Furthermore, it can be seen that the best performance for those two tasks was achieved when we employed MER - KE pair. This observation is in line with Saputra et al. [11] which found that information about medical entities could improve the performance of keyphrase extraction model.

Table 7 Evaluation results of sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE) using pairwise multi-task learning approach. † indicates a stochastically dominant performance over STL baseline. Note that each pairwise MTL model only has scores for the two corresponding tasks, −− represents the score for another task

Model	α	SR	MER	KE
$\text{IndoNLU}_{\text{LARGE}}$ [68] (STL)	−−	92.81	59.59	46.91
Parallel MER - SR	0.3	**92.85**	58.67	−−
	0.5	92.64	58.99	−−
	0.7	92.24	59.43	−−
Hierarchical MER - SR	0.3	92.55	59.80	−−
	0.5	92.52	59.30	−−
	0.7	92.32	59.33	−−
Parallel KE - SR	0.3	92.72	−−	46.36
	0.5	92.70	−−	47.62
	0.7	92.44	−−	48.20
Hierarchical KE - SR	0.3	92.41	−−	46.77
	0.5	92.42	−−	47.72
	0.7	92.66	−−	46.48
Parallel MER - KE	0.3	−−	57.70	49.18
	0.5	−−	58.33	47.61
	0.7	−−	**60.83**	44.77
Hierarchical MER - KE	0.3	−−	57.27	† **49.76**
	0.5	−−	58.68	47.08
	0.7	−−	59.91	45.01

We tested seven different combinations of α and β values for our three-way MTL models. Table 8 presents the results of our three-way MTL experiments. None of the configurations we tested could overperform the best Pairwise MTL model for SR task. While some configurations managed to achieve better performances for MER and KE tasks even when compared to the highest score in pairwise MTL settings, only the parallel model with α value of 0.4 and β value of 0.2 for the KE task that stochastically dominant. We also could not find any specific pattern as we changed the α and β values.

Table 9 shows confusion matrix for SR using Parallel MER - SR model with α value of 0.3, which achieved the best performance among all the models tried in this work. Most of the errors happened because model predicted tokens that were part of question or ignore sentence as part of background sentence. Our further investigation found that the model tends to predict tokens related to

Table 8 Evaluation results of sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE) using three-way multi-task learning approach. † indicates a stochastically dominant performance over STL and Pairwise MTL baselines

Model	α	β	SR	MER	KE
$\text{IndoNLU}_{\text{LARGE}}$ [68] (STL)	−−	−−	92.81	59.59	46.91
Best Pairwise MTL	−−	−−	**92.85**	60.83	49.76
Parallel Three-way	0.33	0.33	92.29	60.16	48.50
	0.40	0.40	92.53	60.26	45.67
	0.40	0.20	92.41	57.73	† **51.28**
	0.20	0.40	86.70	59.10	47.32
	0.60	0.20	92.41	58.66	48.52
	0.20	0.60	86.65	60.35	43.92
	0.20	0.20	86.59	57.91	47.34
Hierarchical Three-way	0.33	0.33	92.47	59.35	48.02
	0.40	0.40	92.13	**61.14**	45.10
	0.40	0.20	92.23	58.94	50.16
	0.20	0.40	92.44	59.87	45.64
	0.60	0.20	92.16	59.91	48.64
	0.20	0.60	89.10	54.60	42.00
	0.20	0.20	92.17	56.89	47.25

Table 9 Confusion matrix for sentence recognition (SR) using Parallel MER - SR model with α value of 0.3

		Predicted		
		Background	**Question**	**Ignore**
Actual	**Background**	10, 090	44	63
	Question	150	2, 683	105
	Ignore	480	20	1, 192

user information (e.g. age, gender, weight, height) as part of background sentence, even though the question asked is not about the user. We noticed that this kind of situation, when a user asked on behalf of others but still included information about himself, appeared several times in the annotated data. Furthermore, we also tried to analyze prediction errors at sentence level. Out of 1,509 sentences in the gold standard, we found that 1,092 (72.37%) of them are exact match with model predictions (i.e. correctly predicted both the boundaries and the type). In other cases, 123 predicted sentences were cut-off in the beginning and/or end, 56 predicted sentences extended into prior or next sentence, and 58 predicted sentences were assigned the incorrect sentence type.

In order to get a better understanding of the MER and KE results, we further analyzed the prediction errors for both tasks. Following Han et al. [75], we identified four types of errors as follows:

- **Missing span (MS)**: Model failed to identify a span as an entity or keyphrase;
- **Unannotated span (US)**: Model identified an unannotated span as an entity or keyphrase;
- **Incorrect span offsets (ISO)**: Model succeeded in identifying a span partially, i.e. the boundary of identified span is incorrect;
- **Incorrect type (IT)**: Model correctly identified a span, but failed to predict the type (e.g. predicted a disease entity as a drug).

For convenience, we only reported the error distributions of a few models with the best performance for each task. Using predictions on the validation set, Table 10 presents prediction errors made by the models on their

Table 10 Distribution of error types for medical entity recognition (MER) and keyphrase extraction (KE). # columns indicate the occurrence number of corresponding error type, while % columns indicate the corresponding ratio

Model	MS		US		ISO		IT	
	#	%	#	%	#	%	#	%
Medical Entity Recognition (MER)								
$\text{IndoNLU}_{\text{LARGE}}$ (STL)	55	11.55	83	17.44	323	67.86	15	3.15
$\text{XLM-R}_{\text{LARGE}}$ (STL)	61	13.23	60	13.02	328	71.15	12	2.60
Parallel MER - KE ($\alpha = 0.7$)	83	17.22	58	12.03	326	67.63	15	3.11
Hierarchical Three-way ($\alpha = 0.4$; $\beta = 0.4$)	69	14.94	69	14.94	309	66.88	15	3.25
Keyphrase Extraction (KE)								
$\text{IndoLEM}_{\text{BASE}}$ (STL)	69	19.06	31	8.56	262	72.38	0	0.00
Parallel MER - KE ($\alpha = 0.3$)	79	22.44	15	4.26	258	73.30	0	0.00
Hierarchical MER - KE ($\alpha = 0.3$)	109	29.46	17	4.59	244	65.95	0	0.00
Parallel Three-way ($\alpha = 0.4$; $\beta = 0.2$)	52	14.99	25	7.20	270	77.81	0	0.00

best run. Note that KE task had no incorrect type errors because it only has one type of span, i.e. keyphrase. It can be seen that most of the prediction errors happened because model failed to determine the boundary of entity or keyphrase correctly, which account for more than 60% of errors.

In order to find out the model ability to generalize when faced with unseen data, we calculated the amount of correctly predicted entities or keyphrases that did not appear in training data. For this purpose, we used predictions on validation data by the best model for each task, i.e. Hierarchical Three-way ($\alpha = 0.4$; $\beta = 0.4$) for MER and Parallel Three-way ($\alpha = 0.4$; $\beta = 0.2$) for KE. We found that the MER model managed to correctly predict 412 unseen medical entities, with 387 of them are unique. Meanwhile, the KE model was able to correctly predict 273 unseen keyphrases, with 258 of them are unique. These findings indicated that both models had a quite good generalization ability. However, further analysis found that the opposite also happened, where both models were sometimes still wrong in predicting entities or keyphrases that already appeared in training data. These incorrect predictions also included common entities and keyphrases. Some medical entities found in this case are: *operasi* (surgery), *demam* (fever), *kista* (cyst), *tumor* (tumor), and *wasir* (hemorrhhoids). Some keyphrases found in this case are: *keputihan* (vaginal discharge), *kemoterapi* (chemotherapy), *pilek* (cold), *sakit kepala* (headache), and *hamil* (pregnant).

Conclusion

Dataset Expansion. In developing a semi-automatic system, one of the most important stages is question processing. For consumer-health system, several modules that can be included in this stage are sentence recognition (SR), medical entity recognition (MER), and keyphrase extraction (KE). This research focuses in two aspects of these three tasks, i.e. dataset and modeling. Since the amount of data used in previous research is relatively small, we conducted annotation on additional data. Using a thousand unlabeled data from Hakim et al. [57], we recruited two annotators for each task and finished the annotation process in three stages. The inter-annotator agreements for SR, MER, and KE tasks were 88.61%, 64.83%, and 35.01% respectively. In the end of annotation process, we have 1, 173 labeled data points that splitted into 773 training instances, 200 validation instances, and 200 testing instances. Additionally, our observation on training data suggested that a sentence that contains medical entities or keyphrases is less likely to be an ignore-type. This finding indicated that information about MER and KE tags has the potential to help the learning process of SR task.

Baseline Establishment. We proposed transformer-based models to solve these three tasks. For single-task learning (STL) settings, we tried 9 different encoder variations and compared the results with CRFs as a baseline. For all tasks, the transformer-based models outperformed the baseline. However, there was no single model consistently achieved the best performance across the three tasks. These results suggested that a larger model did not always achieve a higher score than the smaller ones. There was also no indication which one is generally better between Indonesian and multilingual language models.

Multi-task Learning Models. Based on our earlier observation and Saputra et al. [11] research finding that showed using medical entities as a feature has a positive effect on KE performance, we also tried to implement multi-task learning (MTL) approach. We used $\text{IndoNLU}_{\text{LARGE}}$ as encoder since it obtained the highest average score in STL settings compared to the others. For both pairwise and three-way MTL, we proposed parallel and hierarchical architecture and tried different combinations of loss weights. For each task, there were configurations that managed to achieve a better performance than the STL baseline. In three-way settings, some of the settings managed to achieve a better performance for MER and KE tasks, but none of them could obtain a better score for SR task. These results suggested that additional information regarding the other two tasks could help the learning process for MER and KE tasks, but had only a small effect for SR task.

Appendix A Annotation guidelines

The following are the annotation guidelines used in this research. Rules added for the revision stage are marked with asterisk symbol (*) at the end.

Sentence Recognition (SR)

An inquiry from consumer health forums generally consists of three types of sentences: background, question, and ignore. To determine the type of a given sentence, it is first necessary to identify individual sentences within an inquiry. Since the data in this work are taken from online consumer health forums, separating sentences based on punctuation (e.g. full stop, question mark, exclamation mark) is often not suitable. The following are the rules for determining if a word sequence can be considered as a sentence:

1. A word sequence that follows the rules of proper Indonesian grammar is considered a sentence. This sentence can take the form of declarative sentences with at least a subject and a predicate, an interroga-

tive sentence, an imperative sentence, a greeting, or a request.
2. A word sequence that does not strictly follow proper Indonesian grammar but attempts to be written according to the rules is still considered a sentence, including cases where the sentence does not start with a capital letter, begins with a coordinating conjunction, lacks a subject, or uses improper punctuation.
3. A word sequence in the form of a phrase with a hidden, implicit, or unstated subject or predicate is still considered a sentence.
4. If two consecutive sentences have the same type, they should still be annotated as separate sentences and not merged into a single sentence.*

Medical Entity Recognition (MER)

There are four medical entity types defined in this work: disease, symptom, drug, and treatment. The following aspects should be considered when annotating medical entities in the given text:

1. Words or phrases that are synonyms or abbreviations of another entity are considered two separate entities without including conjunctions, e.g. "*atau*" ("or"), and punctuation marks, e.g. parentheses. For example, both of "*hemoroid*" and "*ambeien*" in the sentence "*pertanda dari beberapa jenis penyakit, seperti hemoroid (ambeien)*" ("a sign of several types of diseases, such as hemorrhoids") should be labeled as two separate disease entities.
2. A misspelling of an entity is still considered an entity.
3. Two distinct entities that are positioned consecutively are still labeled as two separate entities, even if not separated by punctuation.
4. The words "*penyakit*" ("disease") and "*obat*" ("medicine") are only included as part of an entity if the other words combined with them cannot stand alone as an entity. For example, the word "*penyakit*" in the phrase "*penyakit jantung*" ("heart disease") should be included as part of the disease entity, while the disease entity in the phrase "*penyakit diabetes melitus*" ("diabetes mellitus disease") is sufficiently represented by "*diabetes melitus*" ("diabetes mellitus").*

Keyphrase Extraction (KE)

A keyphrase is a word or phrase that provide important information and describe the essence of a document. For inquiries from online consumer health forums, keyphrases can include information about the disease, symptoms experienced, side effects of drugs or health procedures, and other background information that can affect health conditions. The following aspects should be considered when annotating keyphrases in the given text:

1. The number of tokens in a keyphrase is not limited, but aim to keep it as short as possible without losing important information.
2. If a phrase following a conjunction has an independent meaning, it can be treated as a separate keyphrase. However, if it is related to the previous phrase, the phrase after the conjunction is combined with the phrase before and the conjunction itself. For example, phrase "*obat pusing*" ("headache medication") and "*obat demam*" ("fever medication") in the sentence "*saya diberikan obat pusing dan obat demam oleh dokter*" ("I was prescribed headache medication and fever medication by the doctor") should be treated as two separate keyphrase. On the other hand, phrase "*obat batuk dan flu*"("cough and flu medication") should be labeled as a keyphrase as a whole since the word "*flu*" is semantically attached to the word "*obat*".
3. A misspelled important word is still included in the keyphrase.
4. Adverbs providing additional information, such as frequency, e.g. "*sering*" ("often"), and period, e.g. "*sudah seminggu*" ("for a week"), are considered part of the keyphrase.
5. Words that do not add extra information but are placed between two words that form a keyphrase are still included as part of the keyphrase. For example, the word "*saya*" ("my") in the phrase "*kepala saya sering nyeri*" ("my head often hurts") should be a part of keyphrases even though it is not necessarily an important word in this context.
6. Information related to specific conditions that can affect health are treated as keyphrases, e.g. "*hamil*" ("pregnant"), "*lahir prematur*" ("premature birth").*
7. If a keyphrase appears multiple times within the text, all occurrences should be labeled as keyphrases.*

Abbreviations

ASO	Almost Stochastic Order
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CRF	Conditional Random Field
KE	Keyphrase Extraction
LLM	Large Language Model
LSTM	Long Short-Term Memory
MER	Medical Entity Recognition
MTL	Multi-task Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
POS	Part-of-speech
QA	Question Answering

RNN	Recurrent Neural Network
SR	Sentence Recognition
STL	Single-task Learning
SVM	Support Vector Machine

Acknowledgements
This research is supported by Tokopedia-UI AI Center, Faculty of Computer Science Universitas Indonesia

Authors' contributions
T.N., R.M., and A.F.W. design of the work; T.N. did acquisition, analysis, interpretation of data and the creation of new software used in the work. T.N. wrote the first draft of manuscript text; R.M. and A.F.W. wrote and revised manuscript. All authors reviewed the manuscript.

Funding
This work was supported by Faculty of Computer Science, Universitas Indonesia, through Grant NKB- 3/UN2.F11.D/HKP.05.00/2024.

Data availability
Data and code are available at https://github.com/tsqfnfl/sr-mer-ke-idn-chqa.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare the following potential competing interests. Rahmad Mahendra is currently a PhD student at School of Computing Technologies at RMIT University, working at research project at The ARC Training Centre in Cognitive Computing for Medical Technologies. He is also affiliated with School of Computing and Information Systems, the University of Melbourne.

Author details
[1]Faculty of Computer Science, Universitas Indonesia, Kampus UI, 16424 Depok, West Java, Indonesia.

Received: 18 October 2024 Accepted: 8 April 2025
Published online: 06 May 2025

References

1. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, et al. AskHERMES: An online question answering system for complex clinical questions. J Biomed Inform. 2011;44(2):277–88.
2. Weissenborn D, Tsatsaronis G, Schroeder M. Answering factoid questions in the biomedical domain. In: Proceedings of the first Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum (CLEF); 2013 September 27; Valencia, Spain. Aachen: CEUR-WS.org; 2013;1094:1-6. Available from: https://ceur-ws.org/Vol-1094/bioasq2013_submission_5.pdf.
3. Abacha AB, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. Inf Process Manag. 2015;51(5):570–94.
4. Yang Z, Zhou Y, Nyberg E. Learning to Answer Biomedical Questions: OAQA at BioASQ 4B. In: Kakadiaris IA, Paliouras G, Krithara A, editors. Proceedings of the Fourth BioASQ workshop. Berlin: Association for Computational Linguistics; 2016. pp. 23–37. https://doi.org/10.18653/v1/W16-3104.
5. Wiese G, Weissenborn D, Neves M. Neural Domain Adaptation for Biomedical Question Answering. In: Levy R, Specia L, editors. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver: Association for Computational Linguistics; 2017. pp. 281–289. https://doi.org/10.18653/v1/K17-1029.
6. Jin Q, Yuan Z, Xiong G, Yu Q, Ying H, Tan C, et al. Biomedical Question Answering: A Survey of Approaches and Challenges. ACM Comput Surv. 2022;55(2). https://doi.org/10.1145/3490238.
7. Zhang Y, Lu W, Ou W, Zhang G, Zhang X, Cheng J, et al. Chinese medical question answer selection via hybrid models based on CNN and GRU. Multimedia Tools Appl. 2020;79:14751–76.
8. Lamurias A, Sousa D, Couto FM. Generating biomedical question answering corpora from Q &A forums. IEEE Access. 2020;8:161042–51.
9. Ekakristi AS, Mahendra R, Adriani M. Finding Questions in Medical Forum Posts Using Sequence Labeling Approach. In: International Conference on Computational Linguistics and Intelligent Text Processing. Springer; 2018. pp. 62–73.
10. Rohman WN. Pengenalan entitas kesehatan pada forum kesehatan online dengan menggunakan percurrent neural networks [Bachelor's thesis]. Kampus UI Depok: Universitas Indonesia; 2017.
11. Saputra IF, Mahendra R, Wicaksono AF. Keyphrases extraction from user generated contents in healthcare domain using long short-term memory networks. In: Proceedings of the BioNLP 2018 Workshop; 2018 July; Melbourne, Australia. USA: Association for Computational Linguistics; 2018. p. 28–34. Available from: https://doi.org/10.18653/v1/W18-2304.
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics; 2019. pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
13. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
14. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. arXiv:2307.09288.
15. Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, et al. Gpt-ner: Named entity recognition via large language models. 2023. arXiv:2304.10428.
16. Hu Y, Ameer I, Zuo X, Peng X, Zhou Y, Li Z, et al. Zero-shot clinical entity recognition using chatgpt. 2023. arXiv:2303.16416.
17. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, Zhou Y, Li Z, Jiang X, Lu Z, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc. 2024;31(9);1812–20. Available from: https://doi.org/10.1093/jamia/ocad259.
18. Roberts K, Kilicoglu H, Fiszman M, Demner-Fushman D. Decomposing Consumer Health Questions. In: Cohen K, Demner-Fushman D, Ananiadou S, Tsujii Ji, editors. Proceedings of BioNLP 2014. Baltimore: Association for Computational Linguistics; 2014. pp. 29–37. https://doi.org/10.3115/v1/W14-3405.
19. Mahendra R, Hakim AN, Adriani M. Towards question identification from online healthcare consultation forum post in bahasa. In: 2017 International Conference on Asian Language Processing (IALP). 2017. pp. 399–402. https://doi.org/10.1109/IALP.2017.8300627.
20. Abacha AB, Zweigenbaum P. Medical entity recognition: A comparaison of semantic and statistical methods. In: Proceedings of BioNLP 2011 Workshop; 2011 June. Portland: Association for Computational Linguistics; 2011. p. 56–64. Available from: https://aclanthology.org/W11-0207/.
21. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical named entity recognition using deep learning models. In: AMIA annual symposium proceedings. vol. 2017. American Medical Informatics Association; 2017. pp. 1812.
22. Xu K, Zhou Z, Hao T, Liu W. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. In: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. Springer; 2018. pp. 355–365.
23. Cho M, Ha J, Park C, Park S. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. J Biomed Inform. 2020;103: 103381.
24. Yu X, Hu W, Lu S, Sun X, Yuan Z. BioBERT Based Named Entity Recognition in Electronic Medical Record. In: 2019 10th International Conference on Information Technology in Medicine and Education (ITME). 2019. pp. 49–52. https://doi.org/10.1109/ITME.2019.00022.
25. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019;36(4):1234–40. https://doi.org/10.1093/bioinformatics/btz682.

26. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Demner-Fushman D, Cohen KB, Ananiadou S, Tsujii J, editors. Proceedings of the 18th BioNLP Workshop and Shared Task. Florence: Association for Computational Linguistics; 2019. pp. 58–65. https://doi.org/10.18653/v1/W19-5006.
27. Dai Z, Wang X, Ni P, Li Y, Li G, Bai X. Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records. In: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2019. pp. 1–5. https://doi.org/10.1109/CISP-BMEI48845.2019.8965823.
28. Ashok D, Lipton ZC. Promptner: Prompting for named entity recognition. 2023. arXiv:2305.15444.
29. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Rumshisky A, Roberts K, Bethard S, Naumann T, editors. Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis: Association for Computational Linguistics; 2019. pp. 72–78. https://doi.org/10.18653/v1/W19-1909.
30. Suwarningsih W, Supriana I, Purwarianti A. ImNER Indonesian medical named entity recognition. In: 2014 2nd International Conference on Technology, Informatics, Management, Engineering & Environment. IEEE; 2014. pp. 184–188.
31. Sadikin M, Fanany MI, Basaruddin T. A new data representation based on training data characteristics to extract drug name entity in medical text. Comput Intell Neurosci. 2016;2016:3483528. Available from: https://doi.org/10.1155/2016/3483528.
32. Herwando R, Jiwanggi MA, Adriani M, Medical entity recognition using conditional random field (CRF). In: 2017 International Workshop on Big Data and Information Security (IWBIS). IEEE; 2017. pp. 57–62.
33. Hasan KS, Ng V. Automatic keyphrase extraction: A survey of the state of the art. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2014 June Baltimore, Maryland. USA: Association for Computational Linguistics; 2014. p. 1262–73. Available from: https://doi.org/10.3115/v1/P14-1119.
34. Chen M, Sun JT, Zeng HJ, Lam KY. A practical system of keyphrase extraction for web pages. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management; 2005; Bremen, Germany. New YorK: Association for Computing Machinery; 2005. p. 277–8. Available from: https://doi.org/10.1145/1099554.1099625.
35. Nguyen TD, Kan MY. Keyphrase extraction in scientific publications. In: International conference on Asian digital libraries. Springer; 2007. pp. 317–326.
36. Mahata D, Kuriakose J, Shah R, Zimmermann R. Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers); 2018 June; New Orleans, Louisiana. USA: Association for Computational Linguistics; 2018. p. 634–9. Available from: https://doi.org/10.18653/v1/N18-2100.
37. Devika R, Vairavasundaram S, Mahenthar CSJ, Varadarajan V, Kotecha K. A deep learning model based on BERT and sentence transformer for semantic keyphrase extraction on big social data. IEEE Access. 2021;9:165252–61.
38. Dredze M, Wallach HM, Puller D, Pereira F. Generating summary keywords for emails using topics. In: Proceedings of the 13th International Conference on Intelligent User Interfaces; 2008; Gran Canaria, Spain. New York: Association for Computing Machinery; 2018. p. 199–206. Available from: https://doi.org/10.1145/1378773.1378800.
39. Kim SN, Baldwin T. Extracting keywords from multi-party live chats. In: Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation; 2012 November; Bali, Indonesia. Indonesia: Faculty of Computer Science, Universitas Indonesia; 2018. p. 199–208. Available from: https://aclanthology.org/Y12-1021.
40. Papagiannopoulou E, Tsoumakas G. A review of keyphrase extraction. Wiley Interdiscip Rev Data Min Knowl Disc. 2020;10(2):e1339.
41. Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. Inf Sci. 2020;509:257–89.
42. Mihalcea R, Tarau P. Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing; 2004 July. Barcelona, Spain. USA: Association for Computational Linguistics; 2004. p. 404–11. Available from: https://aclanthology.org/W04-3252.
43. Rose S, Engel D, Cramer N, Cowley W. Text mining: applications and theory. Hoboken, USA: John Wiley & Sons, Ltd; 2010. Chapter 1, Automatic keyword extraction from individual documents; p. 1–20.
44. Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG. KEA: Practical automatic keyphrase extraction. In: Proceedings of the Fourth ACM Conference on Digital Libraries; 1999; Berkeley, California, USA. New York: Association for Computing Machinery; 1999. p. 254–5. Available from: https://doi.org/10.1145/313238.313437.
45. Meng R, Zhao S, Han S, He D, Brusilovsky P, Chi Y. Deep Keyphrase Generation. In: Barzilay R, Kan MY, editors. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: Association for Computational Linguistics; 2017. pp. 582–592. https://doi.org/10.18653/v1/P17-1054.
46. Chen J, Zhang X, Wu Y, Yan Z, Li Z. Keyphrase Generation with Correlation Constraints. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, editors. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics; 2018. pp. 4057–4066. https://doi.org/10.18653/v1/D18-1439.
47. Basaldella M, Antolli E, Serra G, Tasso C. Bidirectional lstm recurrent neural network for keyphrase extraction. In: Digital Libraries and Multimedia Archives: 14th Italian Research Conference on Digital Libraries, IRCDL 2018, Udine, Italy, January 25-26, 2018, Proceedings 14. Springer; 2018. pp. 180–187.
48. Kulkarni M, Mahata D, Arora R, Bhowmik R. Learning Rich Representation of Keyphrases from Text. In: Carpuat M, de Marneffe MC, Meza Ruiz IV, editors. Findings of the Association for Computational Linguistics: NAACL 2022. Seattle: Association for Computational Linguistics; 2022. pp. 891–906. https://doi.org/10.18653/v1/2022.findings-naacl.67.
49. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703.
50. Zhu X, Lou Y, Zhao J, Gao W, Deng H. Generative non-autoregressive unsupervised keyphrase extraction with neural topic modeling. Eng Appl Artif Intell. 2023;120:105934. Available from: https://doi.org/10.1016/j.engappai.2023.105934.
51. Chen W, Chan HP, Li P, Bing L, King I. An Integrated Approach for Keyphrase Generation via Exploring the Power of Retrieval and Extraction. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics; 2019. pp. 2846–2856. https://doi.org/10.18653/v1/N19-1292.
52. Wu H, Liu W, Li L, Nie D, Chen T, Zhang F, et al. UniKeyphrase: A Unified Extraction and Generation Framework for Keyphrase Prediction. In: Zong C, Xia F, Li W, Navigli R, editors. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics; 2021. pp. 825–835. https://doi.org/10.18653/v1/2021.findings-acl.73.
53. Cao Y, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. J Biomed Inform. 2010;43(6):962–71. https://doi.org/10.1016/j.jbi.2010.07.007.
54. Sarkar K. A Hybrid Approach to Extract Keyphrases from Medical Documents. Int J Comput Appl. 2013;63(18):14–9. https://doi.org/10.5120/10565-5528.
55. Ding L, Zhang Z, Liu H, Li J, Yu G. Automatic keyphrase extraction from scientific Chinese medical abstracts based on character-level sequence labeling. J Data Inf Sci. 2021;6(3):35–57.
56. Ding L, Zhang Z, Zhao Y. Bert-based chinese medical keyphrase extraction model enhanced with external features. In: International Conference on Asian Digital Libraries. Springer; 2021. pp. 167–176.
57. Hakim AN, Mahendra R, Adriani M, Ekakristi AS. Corpus development for indonesian consumer-health question answering system. In: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE; 2017. pp. 222–227.

58. Tkachenko M, Malyuk M, Holmanyuk A, Liubimov N. Label Studio: Data labeling software. San Francisco: HumanSignal; 2020-2025. Accessed 2023 Sep 18. Available from: https://github.com/heartexlabs/label-studio.
59. Lehman E, DeYoung J, Barzilay R, Wallace BC. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics; 2019. pp. 3705–3717. https://doi.org/10.18653/v1/N19-1371.
60. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
61. Viera AJ, Garrett JM, et al. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360–3.
62. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296–8.
63. Grouin C, Rosset S, Zweigenbaum P, Fort K, Galibert O, Quintard L. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In: Proceedings of the 5th Linguistic Annotation Workshop; 2011 June. Portland: Association for Computational Linguistics; 2011. p. 92–100. Available from: https://aclanthology.org/W11-0411.
64. Deleger L, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, et al. Building gold standard corpora for medical natural language processing tasks. In: AMIA Annual Symposium Proceedings, vol. 2012. American Medical Informatics Association; 2012. p. 144.
65. Brandsen A, Verberne S, Wansleeben M, Lambers K. Creating a dataset for named entity recognition in the archaeology domain. In: Proceedings of the Twelfth Language Resources and Evaluation Conference; 2020 May; Marseille, France. France: European Language Resources Association; 2020. p. 4573–7. Available from: https://aclanthology.org/2020.lrec-1.562.
66. Lafferty J, McCallum A, Pereira F, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Icml, vol. 1. Williamstown; 2001. p. 3.
67. Koto F, Rahimi A, Lau JH, Baldwin T. IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics; 2020. pp. 757–770. https://doi.org/10.18653/v1/2020.coling-main.66.
68. Wilie B, Vincentio K, Winata GI, Cahyawijaya S, Li X, Lim ZY, Soleman S, Mahendra R, Fung P, Bahar S, Purwarianti A. IndoNLU: Benchmark and resources for evaluating indonesian natural language understanding. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing; 2020 December; Suzhou, China. USA: Association for Computational Linguistics; 2020. p. 843-57. Available from: https://doi.org/10.18653/v1/2020.aacl-main.85.
69. Koto F, Lau JH, Baldwin T. IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. In: Moens MF, Huang X, Specia L, Yih SWt, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana: Association for Computational Linguistics; 2021. pp. 10660–10668. https://doi.org/10.18653/v1/2021.emnlp-main.833.
70. Conneau A, Lample G. Cross-lingual language model pretraining. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019. Red Hook, NY, USA: Curran Associates Inc.; 2019. p. 7059-69. Available from: https://doi.org/10.5555/3454287.3454921.
71. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747.
72. Del Barrio E, Cuesta-Albertos JA, Matran C. An optimal transportation approach for assessing almost stochastic order. The Mathematics of the Uncertain: A Tribute to Pedro Gil. Cham, Switzerland: Springer International Publishing; 2018. p. 33–44. (Studies in Systems, Decision and Control; vol 142).
73. Dror R, Shlomov S, Reichart R. Deep Dominance - How to Properly Compare Deep Neural Models. In: Korhonen A, Traum D, Màrquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics; 2019. pp. 2773–2785. https://doi.org/10.18653/v1/P19-1266.
74. Sang EFTK, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL; 2003. Edmonton: Association for Computational Linguistics; 2003. p. 142-7. Available from: https://doi.org/10.3115/1119176.1119195.
75. Han R, Peng T, Yang C, Wang B, Liu L, Wan X. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. 2023. arXiv:2305.14450.

Publisher's Note