AHIAP: An Agile Medical Named Entity Recognition and Relation Extraction Framework Based on Active Learning

Ming Sheng[1], Jing Dong[2(✉)], Yong Zhang[1], Yuelin Bu[3], Anqi Li[4], Weihang Lin[5], Xin Li[6], and Chunxiao Xing[1]

[1] BNRist, DCST, RIIT, Tsinghua University, Beijing 100084, China
{shengming,zhangyong05,xingcx}@tsinghua.edu.cn
[2] University of Queensland, Brisbane, QLD 4072, Australia
j.dong1@uq.net.au
[3] Beijing University of Posts and Telecommunications, Beijing 100876, China
graceyuelin.bu@gmail.com
[4] Beihang University, Beijing 100191, China
anqili99@hotmail.com
[5] Beijing Foreign Studies University, Beijing 100089, China
oscar.lam.bfsu@gmail.com
[6] Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, China
Horsebackdancing@sina.com

Abstract. Knowledge graph plays a significant role in many domains for providing a wide range of assistance. In the medical domain, clinical guidelines, academic papers, Electronic Medical Records (EMRs) and crawled data from the Internet contain essential information. However, those data are usually unstructured but vital to knowledge graph construction. The construction of knowledge graph using unstructured data requires a large number of medical experts to participate in annotations based on their prior experiences and knowledge. Knowledge graphs' quality highly depends on the performances of medical named entity recognition and relation extraction that are both based on data annotation. However, faced with handling such a large amount of enormous data, manual labelling turns out to be a high labor cost task. Besides, the data is generated rapidly, requiring us to annotate and extract quickly to keep the pace with the data accumulation. Therefore, we propose a named entity recognition and relation extraction framework, AHIAP, to solve these problems mentioned above. AHIAP uses active learning method to reduce the labor cost of the annotation process while maintaining the annotation quality. There are two modules in AHIAP, an active learning module for reducing labor cost and a measurement module to control the quality. By using active learning, AHIAP only takes 200 samples to get to the accuracy of 70%, whereas the standard learning strategy takes 4000 records to get the same accuracy.

Keywords: Health knowledge graph · Doctor-in-the-loop · Platform · Healthcare · Health informatics

1 Introduction

Knowledge graphs collect a massive amount of interrelated facts that connect different concepts and instances, and can be transformed into practical knowledge [1]. These linked data triples can be queried by users [2], and support doctors to make diagnostic decisions [3] or develop medical applications while supporting patients to find symptoms-relevant information [4]. Researchers have paid a great amount of effort into the realms of constructing knowledge graphs. There are several developed knowledge graphs available in medical domain like Linked Life Data (LLD), which connecting over 20 bio-databases. However, those health knowledge graphs focus only on storing concept information.

In these health knowledge graphs, RDF triples are stored to represent the knowledge, and there are three types of information stored as nodes: entity, event and concept. We define the knowledge graph with concept nodes as Concept Knowledge Graph (CKG). Correspondingly, we define the knowledge graph with entity nodes and event nodes as Instance Knowledge Graph (IKG). Based on that, we describe the knowledge graph, including both CKG and IKG as Factual Knowledge Graph (FKG) [5]. In the medical domain, IKG contains instance data such as the records of medical, which are useful for further analysis. But these data sources are generally unstructured, in which the knowledge needs to be extracted manually with dramatic labor cost.

To reduce the labor cost, automatic named entity recognition and relation extraction are adopted. Maya Rotmensch et al. [6] presented a methodology for constructing a health knowledge graph using automatic entity extraction from unstructured data. But the mechanical method to do so still requires preprocessed data and a lot of time in model training [7]. Besides, without the doctors to provide useful prior knowledge and measure the process, the quality of the automatic result is relatively unreliable [8].

To solve the labor cost problem, we propose a medical named entity annotation and relation extraction framework AHIAP, which implements active learning to reduce human participation workloads during the medical unstructured data annotation process, and is further combined with "doctor-in-the-loop" methodology [9] to maintain the quality of entity annotation and relation extraction result.

This paper is organized as follows. In Sect. 2 we introduce the related work in the relevant field. In Sect. 3 we present the framework and workflow of AHIAP. In Sect. 4 we show the details of the modules used in this framework, introduce how they reduce the labor cost and perform the quality control. In the end, we summarize the paper and propose future work in Sect. 5.

2 Related Work

In Table 1, nine frameworks that can be used in named entity recognition and relation extraction tasks are listed. They are compared on human participation method and labor cost level. As shown from the table, only three frameworks combine both machine and human effort to accelerate the annotation process with a reliable result. Among all the frameworks listed, only the WebAnno provides full auto annotation but it is only available for project manager and administrators. Most of the named entity recognition and relation extraction frameworks are purely manual.

Table 1. Named entity recognition and relation extraction frameworks

Framework name	Human participation method	Labor cost level
Doccano [10]	Manual	High
BRAT [11]	Manual	High
Prodigy [12]	Semi-auto	Middle
YEDDA [13]	Manual	High
DeepDive:Mindtagger [14]	Manual	High
Anafora [15]	Manual	High
WebAnno [16]	Auto/manual	Low/High
MAE [17]	Manual	High
INCEpTION [18]	Semi-auto	Middle

3   The Framework of AHIAP

The Framework of AHIAP

As Fig. 1 shows, the framework contains three parts: (1) The data source of AHIAP provides unstructured medical data to be annotated; A high-quality health CKG is also an input to provide annotation labels. (2) The building modules. (3) The output of this framework is the high quality annotated medical unstructured data and can be further used to construct high-quality IKG.
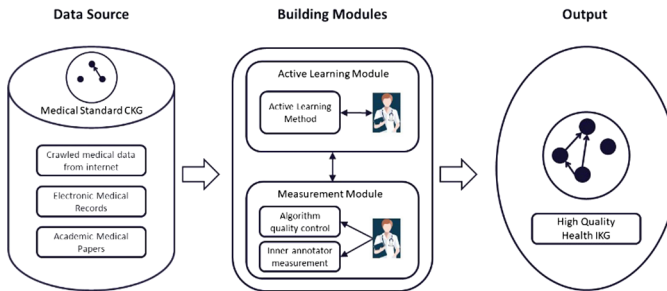
Fig. 1. The framework of AHIAP.

The Workflow of AHIAP

As shown in Fig. 2, in the workflow of AHIAP, the medical unstructured data is taken as input into the active learning module, doctors who are assigned as annotators are asked to annotate a small set of data randomly selected from input dataset. Then, those data are sent to the active learning algorithm to train the model. After initializing a learning model, the algorithm periodically returns the unconfident auto annotation result to the annotator, asks them to prove.

With the model keeping convergence, it becomes more and more accurate and requires less human effort in correction. The measurement module is supervised during the entire process. The fine trained model can automatically extract medical unstructured data and generate high-quality health IKG with almost zero labor cost.
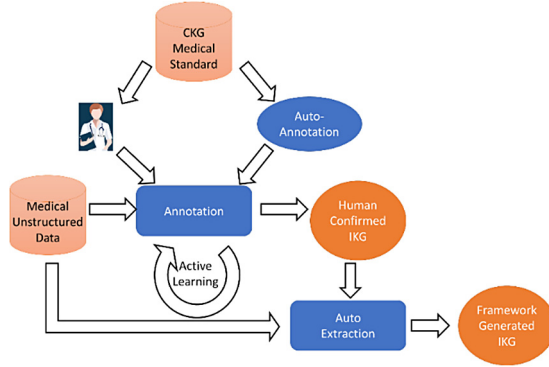

Fig. 2. The detailed workflow of AHIAP to annotate medical unstructured.

4 Modules in AHIAP

In this section, we describe the active learning process of the active learning module in detail, and explain how the mechanism of the measurement module works.

4.1 Deep Active Learning Module

Algorithms that involve humans' interventions can be defined as "human-in-the-loop". Human-in-the-loop has been applied to many aspects of artificial intelligence like named entity recognition [19] and rules learning [20] to improve the performance. Active learning is a machine learning method that involves the human-in-the-loop methodology.

In AHIAP, we use Shen's work [21] to implement an active learning model. When the active learning model is compared with other algorithms, pure deep learning needs a larger labelled dataset to perform well, but when it comes to small datasets, the advantage is less obvious. Meanwhile, expecting better performance with less manually labelling work, active learning methods seek to select a subset of examples that can critically improve the model before ask the annotators to label them.

The deep learning method in our experiment is a CNN-CNN-LSTM architecture including character-level encoder, word-level encoder and tag decoder. The input unstructured data with the low rank will be chosen for active learning use sequence tagging.

As we can see from Fig. 3, using active learning it only takes 200 samples to get to the accuracy over 70%, whereas the standard learning takes more than 4000 records to get the same accuracy. As the number of samples increases, the performance of the model still remains stable. Besides this experiment also shows in the medical field active learning can reduce annotation cost and result in better quality predictors in same time.
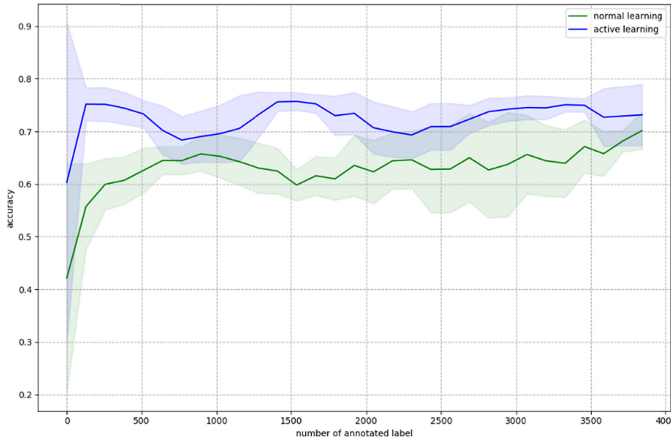
Fig. 3. The converged result between using active learning and standard learning.

4.2 Active Learning Based Named Entity Recognition and Relation Extraction Process

During the learning process, active learning algorithm iteratively queries the most informative instances to manual verification and revision. The appropriate selection of instances in each epoch ensures the cost of manual work being limited in a relatively low level.

Start-up Procedure for Active Learning Process
At the start-up of an annotation assignment, annotation manager initializes it, and determines the field of this assignment; chooses the target medical documents that need to be annotated; a CKG is used for medical standard, and assigned to the doctors. Then, the framework pushes part of randomly selected medical document to doctor, and let them label the data. The labeled data is sent to the measurement module before being transferred to a "storage of training data". The initially trained model is generated through the startup procedure.

Loop Procedure for Active Learning Process
After the startup, there is a loop of the active learning. In this process, the trained machine learning model tries to automatically perform NER on those unstructured medical data not in the training storage, resulting in a machine labelled records. Next, those records are applied for uncertainty-based sampling strategy and calculate the confidence to every machine labelled data. A certain amount of data with the lowest confidence is passed on to the doctors for the annotation. The doctor can choose to accept those annotation results labelled by model or re-annotate them again. Doctors' annotation results are transmitted to the training storage if they can pass the measurement tool. The machine learning module updates based on the new training storage. Finally, the framework starts the next cycle of a loop by applying the trained model to the unstructured medical data

out of the storage. The start-up and loop procedure are where active learning take place and reduce the labor cost.

Termination Procedure for Active Learning Process

The loop terminates once the annotation manager regard that the performance of the model is good enough. The data in training storage and the rest of the machine labelled data is moved to the result storage and becomes the final result of this annotation assignment.

With the growing of dataset to annotate, only a few of data need to be manually processed. Therefore, the labor cost is reduced using this module.

4.3 Measurement Module

In the medical field, system stability is sensitive because they may cause not only some ethical problems but also bring severe medical accidents to patients. Therefore, we must decrease the mistakes of the framework.
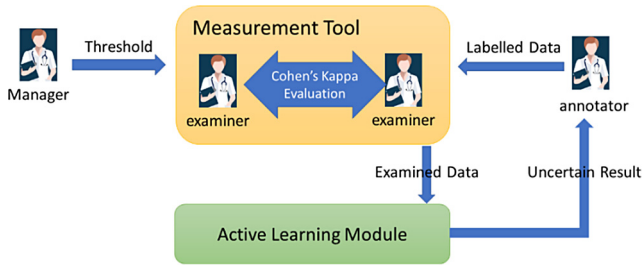

Fig. 4. Workflow of the inner annotator agreement measurement system.

Inner Annotator Agreement Measurement System

To measure the quality of annotated data, the mistakes from the annotator should be minimized. Therefore, we implement an inner annotator agreement measurement system shown as Fig. 4, Cohen's Kappa, has been proved as a very effective agreement measurement evaluation method [22]. We apply that evaluation between the examiners who measure the labelled result from annotator, and only send the examined data which pass the evaluation score threshold to the active learning module. Need to mention that in the framework of AHIAP, examiners, annotation managers and annotators should be doctors with prior knowledge due to the particularity of the medical field.

Up-to-date CKG

This framework using reliable CKG to provide medical standards. During the annotation process, the doctors are asked to choose from CKG standard to annotate on the target corpus rather than self-defined one, which helps doctors to annotate the medical unstructured data and produce reliable and standardized annotation results. Modification to the standards will be strictly restricted.

By applying direct mapping between annotation result and model prediction result with the up-to-date CKG, the high-quality FKG will be generated and ready to use for providing further help in the medical field.

5   Conclusion

In this paper, we propose AHIAP, an agile medical named entity recognition and relations extraction framework used for constructing high-quality health knowledge graph with low labor cost. In AHIAP, we develop two modules to make this framework to require less labor and keep accurate at the same time. The active learning module involves machine learning method with human-in-the-loop mechanism. It makes the trained machine learning model to converge with less data, and eventually, reduces the labor cost in the annotation process. The measurement module ensures the quality of annotation work in real-time by supervising any modification to the data and performing quality control using inner annotator measurement method.

In the future, we will apply AHIAP in the construction of knowledge graphs in more fields for establishing efficient medical support applications, such as cost prediction [23], document analysis [24], entity extraction [25] and recommendation [26]. We are also planning to build a larger CKG with the newest medical standards to improve the performance of this framework.

References

1. Pujara, J., Miao, H., Getoor, L., Cohen, W.: Knowledge graph identification. In: Alani, H., et al. (eds.) ISWC 2013. LNCS, vol. 8218, pp. 542–557. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41335-3_34
2. Verborgh, R., et al.: Triple Pattern Fragments: a low-cost knowledge graph interface for the Web. J. Web Semant. **37**, 184–206 (2016)
3. Donnelly, K.: SNOMED-CT: the advanced terminology and coding system for eHealth. Stud. Health Technol. Inform. **121**, 279 (2006)
4. Agarwala, R., et al.: Database resources of the national center for biotechnology information. Nucleic Acids Res. **45**, D12–D17 (2017)
5. Sheng, M., et al.: DEKGB: an extensible framework for health knowledge graph. In: ICSH, pp. 27–38 (2019)
6. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., Sontag, D.: Learning a health knowledge graph from electronic medical records. Sci. Rep. **7**, 1–11 (2017)
7. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
8. Giorgi, J.M., Bader, G.D., Wren, J.: Towards reliable named entity recognition in the biomedical domain. Bioinformatics **36**, 280–286 (2020)
9. Sheng, M., et al.: DocKG: a knowledge graph framework for health with doctor-in-the-loop. In: Wang, H., Siuly, S., Zhou, R., Martin-Sanchez, F., Zhang, Y., Huang, Z. (eds.) HIS 2019. LNCS, vol. 11837, pp. 3–14. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32962-4_1

10. doccano - Document Annotation Tool. https://doccano.herokuapp.com/. Accessed 11 June 2020
11. brat rapid annotation tool. https://brat.nlplab.org/
12. Prodigy · An annotation tool for AI. Machine Learning & NLP. https://prodi.gy/
13. Jie, Y., Yue Z., Linwei L., Xingxuan L.: YEDDA: a lightweight collaborative text span annotation tool. In: ACL 2018, pp. 31–36 (2018)
14. Deepdive. https://github.com/HazyResearch/deepdive. Accessed 11 June 2020
15. Chen, W., Styler, W.: Anafora: a web-based general purpose annotation tool. In: NAACL, pp. 14–19 (2013)
16. Eckart de Castilho, R., et al.: A web-based tool for the integrated annotation of semantic and syntactic structures. In: LT4DH Workshop, pp. 76–84 (2016)
17. Multi-document Annotation Environment. http://keighrim.github.io/mae-annotation/
18. Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., Gurevych, I.: The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In: ACL, pp. 5–9 (2018)
19. Coelho da Silva, T.L., Magalhães, R.P., et al.: Improving named entity recognition using deep learning with human in the loop. In: EDBT, 594–597 (2019)
20. Yang, Y., Kandogan, E., Li, Y., Sen, P., Lasecki, W.S.: A study on interaction in human-in-the-loop machine learning for text analytics. In: CEUR Workshop (2019)
21. Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. arXiv preprint arXiv:1707.05928 (2017)
22. Vieira, S.M., Kaymak, U., Sousa, J.M.C.: Cohen's kappa coefficient as a performance measure for feature selection. In: WCCI 2010. pp. 1–8. IEEE (2010)
23. Zhao, K., et al.: Modeling patient visit using electronic medical records for cost profile estimation. In: DASFAA, pp. 20–36 (2018)
24. Tian, B., Zhang, Y., Wang, J., Xing, C.: Hierarchical inter-attention network for document classification with multi-task learning. In: IJCAI, pp. 3569–3575 (2019)
25. Wang, J., Lin, C., Li, M., Zaniolo, C.: Boosting approximate dictionary-based entity extraction with synonyms. Inf. Sci. **530**, 1–21 (2020)
26. Zhao, K., et al.: Discovering subsequence patterns for next POI recommendation. In: IJCAI, pp. 3216–3222 (2020)