International Journal of Medical Informatics 106 (2017) 25–31

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Active learning reduces annotation time for clinical concept extraction

Mahnoosh Kholghi[a,b,*], Laurianne Sitbon[a], Guido Zuccon[a], Anthony Nguyen[b]

[a] Queensland University of Technology, Brisbane 4000, Queensland, Australia
[b] The Australian e-Health Research Centre, CSIRO, Brisbane 4029, Queensland, Australia

ARTICLE INFO

Keywords:
Active learning
Annotation time
Machine-assisted pre-annotation
Concept extraction
Clinical free text

ABSTRACT

Objective: To investigate: (1) the annotation time savings by various active learning query strategies compared to supervised learning and a random sampling baseline, and (2) the benefits of active learning-assisted pre-annotations in accelerating the manual annotation process compared to de novo annotation.

Materials and methods: There are 73 and 120 discharge summary reports provided by Beth Israel institute in the train and test sets of the concept extraction task in the i2b2/VA 2010 challenge, respectively. The 73 reports were used in user study experiments for manual annotation. First, all sequences within the 73 reports were manually annotated from scratch. Next, active learning models were built to generate pre-annotations for the sequences selected by a query strategy. The annotation/reviewing time per sequence was recorded. The 120 test reports were used to measure the effectiveness of the active learning models.

Results: When annotating from scratch, active learning reduced the annotation time up to 35% and 28% compared to a fully supervised approach and a random sampling baseline, respectively. Reviewing active learning-assisted pre-annotations resulted in 20% further reduction of the annotation time when compared to de novo annotation.

Discussion: The number of concepts that require manual annotation is a good indicator of the annotation time for various active learning approaches as demonstrated by high correlation between time rate and concept annotation rate.

Conclusion: Active learning has a key role in reducing the time required to manually annotate domain concepts from clinical free text, either when annotating from scratch or reviewing active learning-assisted pre-annotations.

1. Introduction

1.1. Objective

Supervised machine learning (ML) based approaches can effectively extract domain concepts from clinical free texts [1]. However, these approaches are very costly in practice, as they require a large amount of high quality annotated samples to train a powerful ML model. Active learning (AL) is one way to significantly reduce the volume of data requiring manual annotation while not sacrificing the quality of the extracted concepts [2,3]. Most studies on active learning evaluate their approach in simulated settings by considering identical annotation cost per datum. However, actual annotation costs should reflect the time spent on annotation by human annotators. Previous studies in linguistic annotation demonstrated a high degree of variability in annotation time per datum in practice [4–6]. Hachey et al. [4] showed that the samples selected by AL approaches are difficult to annotate (i.e., lower inter-annotator agreement) and take longer (i.e., higher annotation time) compared to other samples. This suggests that the reduction in annotation volumes may be of a different magnitude than actual annotation cost savings.

In this paper, we describe the design of an in-depth user study to measure the actual time required by experts to extract domain concepts from clinical free texts. The recorded annotation times are used to evaluate the impact of a wide range of active learning query strategies on actual time savings compared to supervised learning and a random sampling (RS) baseline. In addition, we investigate how machine-assisted pre-annotations provided by active learning models (i.e., AL-assisted pre-annotations) can further reduce the manual annotation time compared to when annotating from scratch (i.e. de novo annotation). We also study the role of a smart seed selection approach in reducing the annotation time from early batches of active learning. Our previous study demonstrated that Longest Sequence Cluster (LSC) can lead to an initial model with significantly higher effectiveness at early batches of

* Corresponding author at: Queensland University of Technology, Gardens Point Campus, 2 George St, Brisbane, Queensland 4000, Australia.
E-mail addresses: mahnoosh.kholghi@hdr.qut.edu.au (M. Kholghi), laurianne.sitbon@qut.edu.au (L. Sitbon), g.zuccon@qut.edu.au (G. Zuccon), anthony.nguyen@csiro.au (A. Nguyen).

http://dx.doi.org/10.1016/j.ijmedinf.2017.08.001
Received 27 March 2017; Received in revised form 31 July 2017; Accepted 2 August 2017

AL compared to when using RS [7]. We use LSC and RS seed selection approaches to build two initial models. These models are used to pre-annotate the first set of samples selected by an active learning query strategy. The required time to review these two different pre-annotations is then analyzed. We further investigate the effect of smart and random seed set on the overall annotation time reduction.

Our contributions are as follow:

(1) We validate the merits of various active learning query strategies in reducing the burden of manual annotation through a user study. We also study the correlation between the reductions in annotation time and reductions in volume of data that require manual annotation (via simulated measures) across a wide range of AL query strategies.
(2) We use the learning models built across an active learning framework to generate pre-annotations in order to further reduce the manual annotation time. We demonstrate that the time for reviewing these pre-annotations is significantly less than the time spent to annotate samples from scratch. We specifically examine the impact of the pre-annotations' quality in early batches of AL on reducing the annotation time. We show that an initial model built from a "smartly" selected seed set led to higher quality pre-annotations compared to a model built on a randomly selected seed set. This supports the choice of smart seed set to build a strong initial model.

1.2. Background and significance

The high cost incurred by domain experts to manually annotate unstructured text is a major obstacle in efficient information analysis in the clinical domain. Active learning and machine-assisted pre-annotations are two orthogonal approaches to reduce the burden of manual annotation.

Active learning reduces the number of samples (i.e., volume of data) that require manual annotation by selecting a subset of samples that carry more useful information for the model to be built across an AL process [8]. This subset of samples is selected iteratively using approaches called query strategy (QS). A high-performing machine learning model built on the manually annotated subset of samples can be used to automatically generate high quality annotations for the rest of the unlabeled samples. Tomanek and Hahn [6] studied the effect of AL in reducing the annotation time for extracting person, organization, and location entities from the MUC7 corpus. They used the recorded annotation time to compare the performance of only one simple active learning query strategy (i.e., least confidence [9]) with a RS baseline. They found that while there is a high variation in actual annotation time per sample, active learning significantly reduces the actual annotation cost compared to RS by saving 33% of the annotation time. However, the results from a user study conducted to extract clinical concepts (i.e., problem, test, and treatment) showed that there is no significant reduction in annotation time when using AL compared to RS [10]. Chen [10] recruited two annotators to re-annotate a part of the corpus developed for i2b2/VA 2010 concept extraction challenge [11]. Although the simulation results showed that their novel AL query strategy (i.e., CAUSE) [10] reduced the volume of data (i.e., number of the words) that required manual annotation compared to RS, the analysis of the annotation time across two annotators did not correspondingly show reductions in annotation time.

Machine-assisted pre-annotations reduce the manual annotation time by reducing the number of annotations that human annotators must manually add or correct. Pre-annotated data, which is commonly generated using a dictionary [12] or existing NLP systems [13–17], have resulted in significantly reduced annotation time compared to manually annotating the full dataset. Lingren et al. [12] used a dictionary-based approach to pre-annotate disease/disorder and sign/symptom entities with the aim of developing a gold standard for clinical named entity recognition in clinical trial announcements. Fort and Sagot [17] proposed another approach to pre-annotate parts of speech in the Penn Treebank corpus using POS taggers with different levels of accuracy. Both studies found that the time savings by pre-annotations are statistically higher than the de novo annotation. However, the quality of the pre-annotations is directly correlated with savings in annotation time and has an important role in reliability of the output annotations. Inaccurate pre-annotations that require many deletions or corrections, may have an adverse effect by increasing the annotation time [18].

Gobbel et al. [16] investigated the effect of pre-annotations on reducing the annotation time of clinical concepts compared to when annotating from scratch. As opposed to our proposed approach (i.e., to pre-annotate smartly selected samples by active learning models), they randomly selected a set of clinical notes to be pre-annotated using a ML model at each iteration. These pre-annotations were reviewed by human annotators and used to re-train the model. They found that the number of correct pre-annotations was increased when using more reviewed pre-annotations to re-train the model. This subsequently resulted in significant reduction of annotation time by 50%. However, another study by South et al. [19] showed no significant reduction in annotation time when using machine-assisted pre-annotations compared to manual annotation of the full dataset. They used an off-the-shelf de-identification system to pre-annotate a collection of clinical documents and asked seven annotators to review those pre-annotations.

Skeppstedt [20] proposed a possible framework that combines active learning and pre-annotation for Swedish clinical named entity recognition. The idea is to first use active learning models at each iteration to pre-annotate the unlabeled samples. Instead of presenting one annotation for annotators to review, they are asked to choose one of the two most probable annotations generated by an AL model. Therefore, the model has to be confident that one of the two generated pre-annotations is correct. This framework has not been implemented and evaluated.

2. Materials and methods

2.1. Task and data

The annotation task in this study included extracting clinical concepts that belong to one of the three groups of medical problems, tests, and treatments. We followed the concept annotation guideline developed for the concept extraction task of the i2b2/VA 2010 challenge [11]. The i2b2/VA 2010 dataset consists of discharge summaries and progress reports contributed by three institutions (details can be found in Appendix A) and was annotated by 12 clinician and non-clinician annotators. To better understand the implications of the annotation task in practice, it is more appropriate to focus on the data from a single institution.

In our study, we employed the 193 discharge summary reports provided by the Beth Israel Institute, from which 73 and 120 were included in the train and test sets of the task, respectively. The annotated reference provided by the i2b2/VA 2010 challenge [11] for this set of data were used as the gold standard in our user study experiments. The train set, used to conduct the user study experiments, includes 8798 sequences (i.e., roughly corresponds with sentences), 88,722 tokens, and 10,294 target concepts (i.e., 4187 problems, 3035 tests, and 3072 treatments) and comprise more than one third of the total amount of concepts in the i2b2/VA 2010 train set. Table 1 presents additional statistics based on the richness of sequences (i.e., with no concept, with only one concept, and with more than one concept).

2.2. Simulation setup

Simulated experiments include: (1) learning a supervised model from the train set and evaluating it on the test set, and (2) building AL