A Visualization Approach for Rapid Labeling of Clinical Notes for Smoking Status Extraction

Saman Enayati
saman.enayati@temple.edu

Ziyu Yang
tug27634@temple.edu

Benjamin Lu
lu.ben@pennmutual.com

Slobodan Vucetic
vucetic@temple.edu

Abstract

Labeling is typically the most human-intensive step during the development of supervised learning models. In this paper, we propose a simple and easy-to-implement visualization approach that reduces cognitive load and increases the speed of text labeling. The approach is fine-tuned for task of extraction of patient smoking status from clinical notes. The proposed approach consists of the ordering of sentences that mention smoking, centering them at smoking tokens, and annotating to enhance informative parts of the text. Our experiments on clinical notes from the MIMIC-III clinical database demonstrate that our visualization approach enables human annotators to label sentences up to 3 times faster than with a baseline approach.

1 Introduction

Deep learning algorithms achieve state-of-the-art accuracy on a range of natural language processing tasks. However, to achieve high accuracy, deep learning algorithms typically require a lot of labeled data. In extremely error-sensitive applications, such as those in the medical domain, the trade-off between labeling effort and prediction accuracy is strongly skewed towards maximizing the accuracy. In such applications, data labeling arises as the most costly and human-intensive step during the development of deep learning models. In this paper, we focus on a scenario where the requirement is to label all available data because the goal is to maximize the accuracy using the available corpus of documents. In such a scenario, none of the labeling shortcuts developed in the machine learning community such as active learning are of much help on their own.

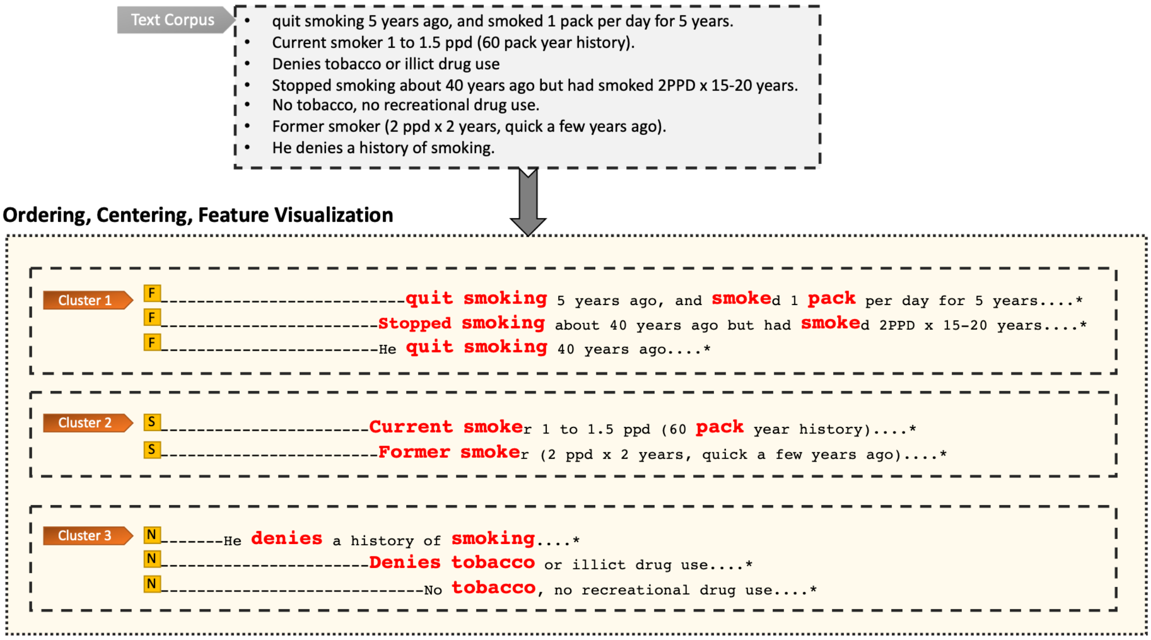
Our focus is on presenting textual information to human annotators in a way that minimizes their cognitive load, thus improving their focus, and maximizes their labeling speed, thus reducing the cost of labeling. Our proposed visualization approach is fine-tuned to enable text labeling in the specific application where the objective is to extract information about smoking status of patients from their medical notes. Smoking status of patients is critical information in many practical applications, ranging from recruiting participants in clinical trials to determining medical and life insurance premiums for prospective customers.

Smoking status extraction is a specific instance of information extraction problems. Our visualization approach relies on several key observations about this particular type of problem. We first observed that smoking status could typically be extracted from sentences that contain one of the smoking keywords such as *smoke, smoking, tobacco, nicotine*. Thus, our first step was to extract from the corpus only sentences containing one of those keywords. Our second observation was that smoking status can typically be deduced from several words surrounding the keyword. Thus, it might be possible to prune very long sentences to sub-sentences surrounding the keyword without loss of information. This observation allows reserving only a single line to display each relevant sentence.

Our third observation is that the space of possible smoking-related sentences occurring in clinical notes is relatively limited and that for any smoking-related sentence there are likely very similar sentences in the corpus. We hypothesized that displaying similar sentences next to each other would allow human annotators to process the text much faster than if sentences are shown in random order. Our fourth observation is that some common discriminative keywords reveal the smoking status, such as *denies, quit, former, packs*. We hypothesized that highlighting those keywords in the text could allow a human annotator to work faster.

Our final observation was that by training a predictive model on the currently available labels, even when the number of available labels is relatively

Proceedings of the 2nd Workshop on Data Science with Human in the Loop: Language Advances, pages 24–30
June 11, 2021. ©2021 Association for Computational Linguistics



Text Corpus

- quit smoking 5 years ago, and smoked 1 pack per day for 5 years.
- Current smoker 1 to 1.5 ppd (60 pack year history).
- Denies tobacco or illict drug use
- Stopped smoking about 40 years ago but had smoked 2PPD x 15-20 years.
- No tobacco, no recreational drug use.
- Former smoker (2 ppd x 2 years, quick a few years ago).
- He denies a history of smoking.

Ordering, Centering, Feature Visualization

Cluster 1
F quit smoking 5 years ago, and smoked 1 pack per day for 5 years....*
F Stopped smoking about 40 years ago but had smoked 2PPD x 15-20 years....*
F He quit smoking 40 years ago....*

Cluster 2
S Current smoker 1 to 1.5 ppd (60 pack year history)....*
S Former smoker (2 ppd x 2 years, quick a few years ago)....*

Cluster 3
N He denies a history of smoking....*
N Denies tobacco or illict drug use....*
N No tobacco, no recreational drug use....*

Figure 1: An illustration of the proposed sequence visualization approach for rapid labeling. The predicted labels for each sentence are shown inside the yellow boxes. where N refers to Non-Smoker, F to Former Smoker, and S to Smoker. Only the 5th sentence in the bottom panel is misclassified by the current prediction model and has to be overwritten by a human annotator.

small, would likely result in prediction accuracy that is significantly higher than a baseline that assigns labels randomly or based on the majority class labels. Thus, providing labels obtained by the current prediction model would allow a human annotator to skip the correctly labeled sentences and only enter the labels for the incorrectly labeled ones. As the number of labels grows, the accuracy of the prediction model is expected to increase, and the effort to correct the labels would decrease, thus increasing the speed of labeling.

The resulting visualization approach developed by exploiting the stated observations is illustrated in Figure 1. A panel at the top shows 7 randomly selected smoking-related sentences from our corpus. A panel at the bottom shows the same sentences displayed using our approach. The main features of our visualization approach are (1) sentence ordering, (2) sentence centering around the smoking keyword, (3) text annotation to emphasize discriminative keywords, and (4) displaying of the predicted labels. We are claiming, and our user study (described in Section 4) confirms it, that the bottom panel makes it much easier and faster for a human annotator to label a large corpus of smoking related sentences for the smoking status of a patient.

To produce the bottom panel in Figure 1, we had to decide (1) what are the smoking keywords, (2) what keywords are discriminative of the smoking status, (3) how to order the sentences, (4) how to provide predicted labels, (5) what to do during the cold start when no or very few sentences are labeled, and (6) how to implement the visualization approach. Details about the proposed approach are provided in Section 3. In Section 2 we provide a brief overview of the related work. In Section 4 we describe the experimental design, explain our user study, and provide experimental results that convincingly indicate the usefulness of the proposed approach.

2 Related Work

Extracting smoking status of patients from Electronic Health Records [EHR] has been crucial in clinical settings, and especially useful to health care providers to select the best care plan for patients at risk of smoking-related diseases. (Rajendran and Topaloglu, 2020) investigates the application of three Deep Learning models on EHR data to extract the smoking status of patients. Authors compare their approach with traditional machine learning models on both binary (Smoker vs Non-Smoker) and multi-class classification (Current Smoker vs. Former Smoker vs. Non-smoker) tasks. (Wang et al., 2016) extracts smoking status from three different sources such as narrative texts, patient-provided-information, and diagnosis codes. They conclude that narrative text proves to be the most useful source for smoking status extraction. (Palmer et al., 2019; Hegde et al., 2018) develop rule-based algorithms to determine tobacco use by patients. (Palmer et al., 2019) further identify the cessation date and smoking intensity of patients. Common for the aforementioned work on smoking status extraction is a need to label sentences and train an appropriate machine learning model. None of those papers discuss issues related to labeling nor attempt to reduce labeling costs.

A common approach to annotate a large amount of data is through crowdsourcing (Fang et al., 2014; Good and Su, 2013; Lim et al., 2020). It has been used in variety of tasks such as Image Classification (Fang et al., 2014), Bioinformatics (Good and Su, 2013), and Text mining (Li et al., 2020). Although crowdsourcing is a cost-effective way to collect labeled data, it can still be costly when the required labeling effort is significant. Moreover, when using imperfect annotators with varying levels of expertise, it is important to develop appropriate label integration approaches (Settles, 2011). Beyond the crowdsourcing issues, one popular approach to reduce labeling costs is to apply Active Learning and label only the most informative examples (Fang et al., 2014).

More recently, Human-In-the-Loop [HIL] approaches were proposed to improve the efficiency of annotation (Klie et al., 2020; Kim and Pardo, 2018). (Kim and Pardo, 2018) present a HIL system for sound event detection, which directs the annotator's attention to the most promising regions of an audio clip for labeling. (Klie et al., 2020) apply a similar technique on Entity Linking [EL] task, in which the machine learning component makes recommendations about the most relevant entries in a knowledge base, and the annotator selects the correct candidate. The recommender improves itself based on the obtained feedback. In addition, (Qian et al., 2020) present an interface for entity normalization annotation in which they measure the number of clicks in a tool to quantify the human effort.

While many papers attempt to minimize labeling effort, a vast majority of them are measuring the effort by counting the number of labeled examples. There are very few papers (Zhang et al., 2019) that measure labeling effort in terms of elapsed time. The uniqueness of our work is in demonstrating that annotation speed can be significantly impacted by the way data is presented to an annotator. Furthermore, our work is specific in its focus on an extreme labeling scenario where the task is to label the complete corpus in order to maximize the prediction accuracy.

3 Methodology

Problem Definition: Given a document corpus D representing clinical notes of patients from which a set of N unlabeled smoking-related sentences $S_1, S_2, ..., S_N$ is extracted, the goal is to ask human annotators to label all N sentences for smoking status. There are 4 types of labels: *Smoker (S), Non-Smoker (N), Former Smoker (F), and Other (O)*, where *Others* refer to sentences that do not reveal the smoking status.

In this section, we describe a visualization approach that improves human annotation speed. The main components of the approach are sequence ordering, label prediction, and text visualization. The details are explained in the following subsections.

3.1 Ordering

Our goal is to order sentences in a computationally-efficient manner by combining clustering and alignment algorithms. We use clustering to find groups of similar sequences that will subsequently be ordered with help of an alignment algorithm.

In order to cluster sentences, we rely on their vector embeddings. In particular, we use sequence embeddings of the pre-trained BERT model (Devlin et al., 2019). K-Means Clustering, whose computational cost is $O(N)$ as implemented by (Pedregosa et al., 2011), is used to find k clusters, where k is selected such that the average cluster size is limited to a specified size.

Sentences in each cluster are then ordered, such that neighboring sentences are perceived by a human annotator to be as similar as possible. Rather than ordering sentences based on BERT embeddings, we instead resort to sequence alignment distance, which we hypothesize are closer to human perception of similarity. In particular, we apply Needleman–Wunsch algorithm [NWA] [1] (Needleman and Wunsch, 1970), which is a dynamic programming algorithm that finds a similarity score between a pair of sentences in $O(L^2)$ time, where L

[1] http://emboss.sourceforge.net/docs/emboss_tutorial/node3.html

is the length of a sentence, For each cluster, we create a pairwise score matrix, $Score$, of size $N_c \times N_c$, where N_c is the number of sequences within the cluster c.

To find the order of the sentences in each cluster, we apply the following greedy algorithm. It starts by selecting the first sentence at random. The next sentence is its nearest neighbor, according to $Score$ matrix. The process continues by adding the nearest neighbors of previous sentences.

3.2 Sentence Visualization

Once the sentences are sorted, our next objective is to display them in a way that reduces the cognitive load of a human annotator. Our first idea is to center the sequences around smoking-related keywords such as *Smoke, Smoking, Tobacco, Nicotine*. We find those keywords by applying word2vec (Mikolov et al., 2013) to our document corpus D and by finding neighbors of word *Smoke* in the resulting embedding. Then, we manually select neighbors that are indicative of smoking-related sentences.

According to the maximum screen width, we align the sentences such that the smoking keyword appears in the middle of the screen. In addition, we fill the empty spaces before the sentence starts with dashes (-) to improve readability.

Our labeling approach proceeds in batches. After selecting the first batch of M unlabeled sentences at random (in our experiments we use $M = 200$), we do not display any predicted labels and orders. After we obtain labels for the first batch, we train a baseline machine learning model such as logistic regression using the bag of words representation (in our experiments we used the most frequent 500 non-stop words). Then, we analyze the statistical significance of the logistic regression weights and select K words associated with the most significant weights as discriminative words. Examples of discriminative keywords are *cigarette, denies, quit, former, packs*.

We select the second batch of unlabeled sentences at random, order them, and display them centered with the discriminative words in bold red font to improve readability. In addition, we display the predicted labels by the logistic regression next to the ordered sentences.

Rather than building a specialized sentence visualization and annotation tool, we use MS Excel[2]. Each sentence occupies one row in the Excel spreadsheet, where the first column is reserved for prediction labels, and the second column is reserved for the centered annotated sentences. An advantage of Excel is that it enables the use of the built-in cell drag feature to quickly change annotations of neighboring sentences. In addition, we use *Courier* as the font format, since it is a monospaced font type. The monospaced font displays each character or letter in the same amount of horizontal space. As a result, it makes the alignment and centering precise.

We continue selecting batches, labeling them, and retraining the prediction models. Once the number of labels becomes sufficiently large ($1,000$ in our experiments) we replace logistic regression with deep learning. We also allow for the batches to become larger over time.

4 Experimental Design

We performed our experiments using 52,726 discharge notes from the MIMIC-III dataset (Johnson et al., 2016), which contains de-identified records of the Beth Israel Deaconess Medical Center's Intensive Unit emergency department patients from 2001 to 2012.

We defined smoking-related keywords by selecting keyword *smoke* and its selected word2vec nearest neighbors. We collected 26 unique keywords. Using those keywords, we found 34,149 unique matching sentences.

4.1 Results

We evaluate the effectiveness of our proposed approach in three different rounds of labeling. We performed a user study with 2 human annotators (the first two co-authors of this paper) to measure labeling time in each of the 3 rounds of labeling. The total number of sentences annotated by each user in our experiments was 3,000 sentences each. In addition, in Section 4.2, we performed an ablation study to analyze the impact of different components of the proposed visualization approach.

In addition to labeling time, we also report the labeling rate, which is the number of sentences labeled per minute:

$$Rate = \frac{\textit{\# of annotated sequences}}{\textit{elapsed time}}$$

[2] https://www.microsoft.com/en-us/microsoft-365/excel

Groups & Settings	User 1 (mins)	User 2 (mins)	Rate User 1 (Sent/min)	Rate User 2 (Sent/min)	Total rate (Sent/min)
Round 1					
Batch1 (Unordered)	27	19	7	10	17
Round 2					
Batch1 (Unordered)	19	17	10	11	21
Batch2 (Ordered)	12	11	16	17	33
Batch3 (Ordered)	**11**	**9**	**17**	**21**	**38**
Batch4 (Unordered)	16	16	12	12	24

Table 1: The annotation results in Round 1 and 2. The experiments are conducted in the same order as the numbers indicate. Each group contains 200 sentences. Unordered refers to the baseline, and Ordered is our visualization approach.

Groups and Settings	User 1 (mins)	User 2 (mins)	Rate User 1 (Sent/min)	Rate User 2 (Sent/min)	Total rate (Sent/min)
Batch1 (Unordered)	40	35	12	14	26
Batch2 (Ordered)	23	23	21	21	42
Batch3 (Ordered)	**19**	**20**	**26**	**24**	**50**
Batch4 (Unordered)	34	34	14	14	28

Table 2: The results for Round 3. The experiments are conducted in the same order as the numbers indicate. Each Group contains 500 samples. The labels for these experiments are provided by fine-tuned Clinical BERT model. Unordered refers to the baseline, and Ordered is our visualization approach.

In the following subsections, we explain the basics of each baseline method as well as the experimental design for each round of labeling.

4.1.1 Round 1

In this round of the experiment, we select 200 random sentences. We display them in the same way as it is shown in the upper panel in Figure 1. Once we obtain the labels from the first batch, we train a logistic regression model. The first row of Table 1 shows the annotation details.

4.2 Round 2

We asked users to annotate 800 sentences in 4 batches. We chose the Latin square design to proceed as unordered, ordered, ordered, and unordered batches. We have also use logistic regression model to predict the labels for all the batches. Table 1 demonstrates the result of this round.

On average, the annotation rate using our method is 1.9× compared to round 1. Additionally, it is 1.5× faster compared to the unordered set in Round 2. By repeating the annotation task in batches 3 and 4, we can speed up the rate in our method by 15% (from 33 to 38) and in the unordered set by 14% (from 21 to 24).

4.2.1 Round 3

We annotated 2,000 sentences in 4 batches, each batch containing 500 sentences. Similar to Round 2, we set up the experiments with the Latin Triangle mixture design (unordered, ordered, ordered, unordered).

Given the annotated data from Round 1 and 2, we replaced the classifier with a deep learning algorithm. We use the Clinical BERT, which is pretrained on all the discharge summary notes in the MIMIC dataset. We split the data into 800 training and 200 for testing. The hyperparameters are selected according to (Devlin et al., 2019). We set the batch size to 16, learning rate to $2e{-}5$, maximum sentence length to 200, and fine-tuned it for 4 epochs. We have also performed experiments with SVM, logistic regression. Table 3 demonstrates the performance of all the classifiers.

According to Table 2, the annotation rate increased from Round 2 to Round 3 by 29% (from 35.5 to 46) with our approach. However, it increased by 16% (from 22.5 in Round 2 to 27 in Round 3) using the baseline approach.

Comparing the annotation speed in Round 3, our approach is 1.7× faster than the baseline (46 compared to 27). Since the size of the batches increased in Round 3, there was more redundancy in the sentences and our approach was more helpful to the annotators than in Round 2. In particular, ordering resulted in smoother transitions between sentences, which contributed to faster human annotation.

Last but not the least, by repeating the labeling task, we expect users to get used to the data, and therefore, we expected the annotation rate to increase regardless of the visualization approach. Confirming this assumption, users on average got 19% faster with our method during Round 3 (rate increased from 42 to 50), while they got only 7% faster with the baseline approach (rate increased from 26 to 28).

Model	**Accuracy Round 1**	**Accuracy Round 2**
Baseline	0.35	0.36
Logistic Regression	0.76	0.79
SVM	0.78	0.80
Fine-tuned Clinical BERT	0.78	0.89

Table 3: All the classifiers are trained to predict 4 classes: Smoker, NonSmoker, Former, and Other. Baseline accuracy is the fraction of the majority class in the test set. In Round 1, there are 800 training and 200 test sentences. In Round 2, there are 3,400 training and 600 test sentences.

4.3 Ablation Study

In this section, we analyze the impact of two components of our system on the final annotation rate. We asked one of the users to annotate an additional 1,000 sentences. We split the set into two groups, each group with 500 samples. First, we studied the impact of centering. Therefore, we aligned all the data to the left and kept the ordering and feature visualization. Second, we removed the feature visualization component, and kept the ordering and centering. Table 4 shows the results of these two experiments.

Components	**User 2 (mins)**	**Rate User 2 (Sent/min)**
No centering	22	22
No coloring	21	23

Table 4: Ablation study on the impact of centering and feature visualization. In the first row, we do not center the sentences around the smoke keywords. In the second row, we do not highlight the important features.

According to the results for Round 2 in Table 2, the highest rate for User 2 was 24 sentences per minute. However, when we removed the centering component, the rate decreased by 8%, to 22 per minute. In addition, by removing the coloring component, the rate decreased by 4%, to 23 per minute. The centering component had a stronger impact on the labeling rate than the coloring component. However, both of the removals reduced the rate of labeling.

Given the annotated data from the ablation study, and adding all the labeled data from the first and second rounds, we re-trained all the classifiers on 3,400 training sentences and used 600 sentences for testing. We observed 15% improvement in the BERT model accuracy and 3% improvement in the Logistic Regression model accuracy compared to the models trained on Round data.

5 Conclusion

We presented a visualization approach that enables rapid annotation of sentences for smoking status of patients. Our framework contains three main components: sentence ordering, sentence presentation, and sentence labeling by the prediction model. Our approach does not depend on high-quality ML predictors to provide initial labels. The display has a significant impact on speeding up the annotation process. We evaluated our visualization approach with a user study on sentences from MIMIC-III discharge summaries. We achieved close to 3× faster annotation rate compared to the baseline method that displayed sentences randomly in their original shape. As the annotation progressed, as the batches of unlabeled sentences became larger, and as the prediction models improved, the annotation speed kept increasing in our user experiments. The proposed visualization approach is applicable to similar text classification tasks. It is a topic of further research to study how to modify the presented approach to make it applicable to a large number of text annotation tasks in natural language processing.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Meng Fang, Jie Yin, and Dacheng Tao. 2014. Active learning for crowdsourcing using knowledge transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Benjamin M. Good and Andrew I. Su. 2013. Crowdsourcing for bioinformatics. *Bioinformatics*, 29(16):1925–1933.

Harshad Hegde, Neel Shimpi, Ingrid Glurich, and Amit Acharya. 2018. Tobacco use status from clinical notes using natural language processing and rule based algorithm. *Technology and Health Care*, 26(3):445–456.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Bongjun Kim and Bryan Pardo. 2018. A human-in-the-loop system for sound event detection and annotation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–23.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2020. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6982–6993, Online. Association for Computational Linguistics.

Maolin Li, Hiroya Takamura, and Sophia Ananiadou. 2020. A neural model for aggregating coreference annotation in crowdsourcing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5760–5773, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Ellen L Palmer, Saeed Hassanpour, John Higgins, Jennifer A Doherty, and Tracy Onega. 2019. Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. *BMC medical informatics and decision making*, 19(1):1–10.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kun Qian, Lucian Popa, and Yunyao Li. 2020. An intuitive user interface for human-in-the-loop entity name parsing and entity variant generation. In *Proceedings of (DaSH@KDD)*. Association for Computing Machinery.

Suraj Rajendran and Umit Topaloglu. 2020. Extracting smoking status from electronic health records using nlp and deep learning. *AMIA Summits on Translational Science Proceedings*, 2020:507.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Liwei Wang, Xiaoyang Ruan, Ping Yang, and Hongfang Liu. 2016. Comparison of three information sources for smoking information in electronic health records. *Cancer informatics*, 15:CIN–S40604.

Shanshan Zhang, Lihong He, Eduard Dragut, and Slobodan Vucetic. 2019. How to invest my time: Lessons from human-in-the-loop entity extraction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, page 2305–2313, New York, NY, USA. Association for Computing Machinery.