Weakly Supervised Medical Entity Extraction and Linking for Chief Complaints

Anonymous ACL submission

Abstract

A Chief complaint (CC) is the reason for the medical visit as stated in the patient's own words. It helps medical professionals to quickly understand a patient's situation, and also serves as a short summary for medical text mining. However, chief complaint records often take a variety of entering methods, resulting in a wide variation of medical notations, which makes it difficult to standardize across different medical institutions for record keeping or text mining. In this study, we propose a weakly supervised method to automatically extract and link entities in chief complaints in the absence of human annotation. We first adopt a split-and-match algorithm to produce weak annotations, including entity mention spans and class labels, on 1.2 million real-world de-identified and IRB approved chief complaint records. Then we train a BERT-based model with generated weak labels to locate entity mentions in chief complaint text and link them to a pre-defined ontology. We conducted extensive experiments and the results showed that our Weakly Supervised Entity Extraction and Linking (WESEEL) method produced superior performance over previous methods without any human annotation.

1 Introduction

A chief complaint (CC) is an initial statement of patient derived medical issues, which is often elicited prior to formal medical tests and diagnoses. It provides a brief statement about a patient's reasons for encounter, current symptoms, and medical history (Chang et al., 2020). It can be of great help to medical professionals in understanding a patient's situation and lead to the appropriate diagnoses and treatments (Wagner et al., 2006). Furthermore, it can be seen as a summary of patient profiles and medical records, and it is useful for a wide variety of medical text mining tasks. But the application of chief complaints is greatly restricted by the fact that there lacks a widely accepted standard for data entry of chief complaints. Hospitals and health care systems adopt different ontology and standards to enter and store chief complaint data (Horng et al., 2019), which causes chief complaint records contain various local terminologies and abbreviations.

Commonly, a chief complaint record is a piece of free-text statement and may contain one or multiple medical entity mentions. In table 1, several chief complaint records and the corresponding concept annotations in HaPPy ontology (Horng et al., 2019) are shown, which also stress the typical difficulties in understanding chief complaints. First of all, synonyms are very common in chief complaints. The same concept can appear in different forms, e.g. "chest pain", "CP", or "cerebral palsy". Another common issue is word sharing. For example, in the record of "migraine with neck/back pain, fever", the word "pain" is shared by both "neck pain" and "back pain". Also, in a more complex example "chills/body aches/n/d/r side jaw pain", the slash "/" acts not only as the separator to split multiple entities, but also a part of the abbreviation "n/d" (i.e. nausea and diarrhea). In these cases, it is even challenging for domain experts to divide and identify entities. Moreover, a chief complaint record is often entered as a piece of free text, where misspellings can occur commonly. All these examples exhibit the challenges in processing chief complaints with automatic NLP techniques.

Due to the difference in terms of data format (long vs. short), target applications (general purpose vs. emergency) and concept ontology (the number of concepts general medical records is much larger), many tools designed for general medical domain cannot be directly applied to CC documents. Several attempts have been made to automate the process of mining entities in chief complaints. For example, Karagounis et al. utilized rule-based matching tool MetaMap (Aronson, 2001) to map free-text chief complaints to ICD-10-CM codes. Chang et al. fine-tuned a BERT model

Chief complaint Records	Concepts in Ontology	NLP Challenges
urinary tract pain	dysuria	Synonym understanding
migraine with neck/back pain, fever	migraine, neck pain, back pain, fever	Resolving shared tokens
chills/body aches/n/d/r side jaw pain	chills, body aches, nausea, diarrhea, jaw pain	Segmentation & abbreviations
ha light headed fatigue r arm pain	headache, dizziness, fatigue, arm pain	Missing separators

Table 1: Examples of chief complaint records and corresponding concepts in HaPPy ontology.

on 1.8 million emergency department (ED) chief complaint records to derive a domain-specific representation, which achieved improved performance on predicting chief complaint concepts. Despite these advancements, most existing efforts (Dara et al., 2008; Conway et al., 2013; Duangsuwan and Saeku, 2018; Lee et al., 2019; Valmianski et al., 2019; Hsu et al., 2020; Osborne et al., 2020) on automatic chief complaint processing viewed the task as a classification problem, whereas in the real-world case, one chief complaint record often contains multiple entities (e.g. the examples shown in Table 1). Consequently, it is more appropriate and useful to view the task as entity extraction problem (i.e., identifying the actual mention span of each entity in a free text) and entity linking problem (i.e., linking each entity mention to an entry in a chief complaint ontology).

Besides, nowadays most machine learning applications require extensive annotated data to train their models (Hsu et al., 2016; Liao et al., 2015; Rochefort et al., 2015; Shen et al., 2017). However, in the case of medical domain, the data annotation is prohibitively sensitive and expensive, due to the data privacy concern and high cost of recruiting healthcare professionals. The lack of high-quality annotation also hinders the progress of NLP methods and applications for chief complains.

In this study, we aim to advance the application of NLP to chief complaints by proposing a novel task setup consisting of two steps: extracting entity mentions from a chief complaint record and linking them to a given ontology. We utilize a BERT-based extraction and linking model to address this task and propose a split-and-match (S&M) algorithm to generate weak annotations to resolve the shortage of annotated data. We conducted experiments with 1.2 million free-text ED chief complaint records from local hospitals and the results show that the proposed method can achieve better performance comparing with various baselines.

Specifically, the contributions of our work are:

- This is the first study mining entities in chief complaints with two explicit steps (extraction and linking), which is more advantageous than the classification setup in previous studies.
- We propose a weak supervision method WESEEL for extracting and linking entities in chief complaints. We demonstrate the superiority of our method over various baselines with extensive experiments and analyses.
- We contribute a new dataset, containing 1.2 million free-text chief complaint records from emergency departments of local hospitals, and 1,013 data examples are manually annotated by clinicians for the purpose of testing.

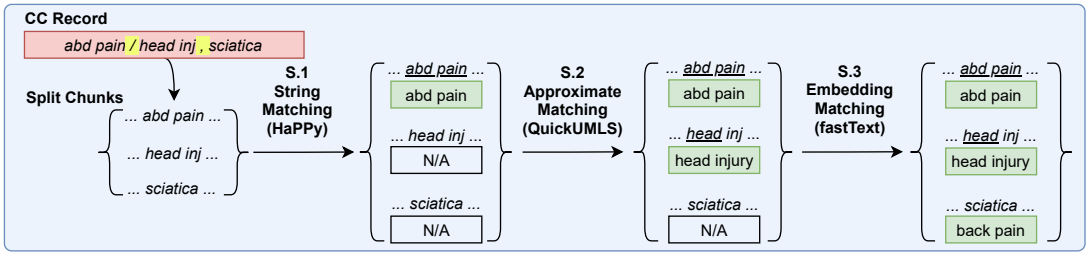
2 Related Works

2.1 Studies on Chief Complaint

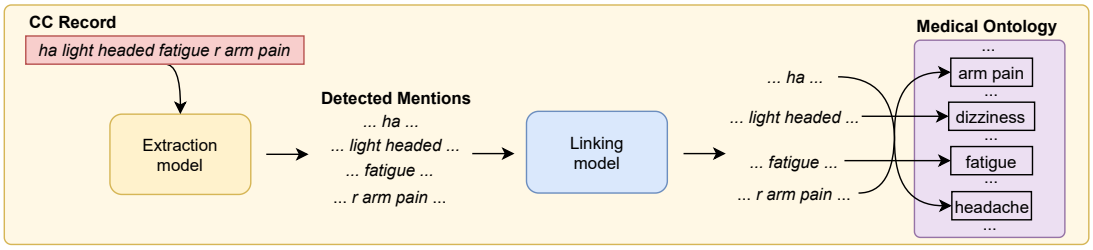
Previous works had adopted automatic chief complaint processing in various medical tasks. Chief complaint records were utilized for syndromic surveillance (Travers et al., 2007; Dara et al., 2008; Conway et al., 2013). Chief complaint records were often studied for specific medical issues. Based on chief complaint records, Devlin et al. (2019a) identified the mentions of Gout flares, Hsu et al. (2020) classified the mentions of inflenza-like illness, and Fernandes et al. (2020) predicted Intensive Care Unit (ICU) admission.

Recently, single-label classification was performed on 2.1 million patient-level ED visit records with ICD codes (Lee et al., 2019). Valmianski et al. (2019) compared embedding from BERT and its variants BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019) with TF-IDF on a dataset of 200,000 patient-generated reasons-for-visit entries and mapped them to 795 unique chief complaint concepts. Teng et al. (2020) formulated automatic diagnose code assignment as multi-label classification to predict ICD codes from free-text medical records including chief complaints. Most datasets in these studies are not publicly available.

Despite several attempts over the years to summarize chief complaints in a standard way (Travers and Haas, 2006; Haas et al., 2008; Aronsky et al.,



(a) Process flow of weak label generation. Three examples are shown and successfully matched to concepts in the ontology at different stages (indicated in green box).



(b) Schematic diagram of our model for entity extraction and linking in chief complaint.

Figure 1: Overview of our proposed method WESEEL.

2001), it is still a serious obstacle that different health care systems have different ways of summarizing chief complaint records, which greatly hampered the usefulness of chief complaint records to downstream medical NLP tasks (Horng et al., 2019). In 2019, Horng et al. (2019) derived an ontology for chief complaints, called Hierarchical Presenting Problem ontology (HaPPy), containing 692 chief complaint concepts and successfully captured 95.9% of 180,424 consecutive ED patient visit records. This was the first publicly available chief complaint ontology, and can be used as ground-truth labels to predict concepts. Nevertheless, the string matching based methods have very low recall since chief complaint recording is poorly standardized in practice.

2.2 Medical Entity Extraction and Linking

Named Entity Recognition (NER) is a common NLP technique for locating entity mentions in a text. Many studies have been conducted to examine the entities in medical text and develop automatic methods to extract them. Zhao et al. (2019) developed a multi-task framework to jointly tackle the task of medical NER and entity normalization. Bhatia et al. (2019) developed a web service for medical NER and relationship extraction in medical data. To tackle the issue of limited data annotation, Hofer et al. (2018) proposed a method for few-shot learning of NER in medical domain.

Entity linking (EL) aims to map entity mentions in text to concepts in a knowledge base or ontology. String matching- and rule-based systems like cTAKES (Savova et al., 2010) and MetaMap (Aronson, 2001) were proposed to link medical entities in EHRs. Recent studies also utilized neural networks. MedType (Vashishth et al., 2020) was pretrained with large auto-annotated datasets such as WikiMed, PubMedDS and EHR documents, and achieved state-of-the-art performance on multiple medical entity linking benchmarks. Chen et al. (2021) applied BERT to learn alignment between mentions and entity names. However, most of those studies requires a fixed list of mention candidates, which does not apply to our task setting.

2.3 Resolving Shortage of Annotated Data

Due to the above constraints, the supervision paradigm to deal with chief complaint task falls under weak supervision, which is closely related to few-shot learning (Yang and Katiyar, 2020) and interactive learning (Greenbaum et al., 2019). Fidelity Weighted Learning (Dehghani et al., 2017a) uses a teacher-student model, in which the teacher model has access to the annotated data and the student model is supervised by the output of the teacher. Hedderich and Klakow (2018) proposed to add a noise adaptation layer to LSTM to correct noisy labels in training data. All of these learning paradigms need a small amount of annotated data to ramp up the training, which involves expensive expert annotation and is difficult to scale up.

Weakly supervised method is effective in training machine learning models with automatically generated pseudo-labeled data. It has been utilized in sentiment analysis (Severyn and Moschitti,

2015), relation extraction (Hoffmann et al., 2011; Bing et al., 2015), and information retrieval (Dehghani et al., 2017b). It is also used in medical topics such as drug-drug interactions (Li et al., 2016), medical term identification (Névéol et al., 2017), and sentence extraction in clinical trial reports (Wallace et al., 2016).

Original Records	After Expansion
hand, knee pain	hand pain, knee pain
head and elbow injury	head injury and elbow injury
pain in r side of face, eye	pain in r side of face, pain in eye

Table 2: Examples of chief complaint records with word sharing and the corresponding expanded texts.

3 Methodology

Formally, chief complaint entity extraction and linking can formulate as follows. A set of medical concepts in an ontology is denoted as $C = \{c_1, c_2, ..., c_N\}$ and a free-text chief complaint record is a sequence of words $D = (w_1, w_2, ..., w_{|D|})$. The goal of the task is to (1) identify one or multiple entity mentions $M = \{m_1, m_2, ..., m_{|M|}\}$, each indicated by an index span $(w_i, ..., w_j)$ of D, and (2) link them to corresponding concepts $c \in C$ in the ontology.

To reduce the reliance of human annotation, we utilize a split-and-match algorithm to generate weak training labels for both extraction and linking. We build a sequence labeling model with BERT (CCME-BERT) as the extraction model to identify chief complaint entity mentions. Then a BiLSTM model (CCEL) is adopted as the entity linking model to combine the word-level and character-level embedding of entity mentions, followed by a feedforward neural network to predict concept label in an ontology. The schematic diagram of the proposed label generation process and the entity recognition model is shown in Figure 1.

3.0 Pre-processing Chief Complaint Records

We observe that a large percent of chief complaint records (55.7% on labeled subset) contain multiple symptoms or reasons, and triage nurses often use punctuation marks to delimit multiple parts. Thus we manually select 10 common punctuation marks that are used as separators.

Besides, many chief complaint records have word sharing issues that occurs between two entity mentions, which often talk about body parts. For example, the word "pain" in "hand, knee pain" is shared by both "hand pain" and "knee pain". This issue can be troublesome for models to predict contiguous spans for entity mentions. Thus, for each record, we detect whether a word is shared by multiple entities with the help of THBP, an ontology for normalizing names of human body parts (Wang et al., 2019), then expand the record by inserting the shared word at appropriate places. Table 2 shows examples of mentions with word sharing and the corrected chief complaint records after expansion.

3.1 Generating Weak Labels for Training

Weak supervision trains machine learning models with noisy sources to provide supervision signals. It is common when human annotated data is limited or unavailable. The key for a successful weak supervision is to derive effective pseudo labels that carry the inductive bias of the target task from unannotated data. In our case, a large proportion of chief complaint records contain explicit punctuation separators that can segment a chief complaint record into multiple chunks and some of them can be chief complaint entity mentions. Thus, we propose a split-and-match algorithm to automatically generate weak labels of chief complaint entity mentions and concept labels.

1. **Split**: We split a chief complaint record into multiple chunks by pre-defined separators.

2. **Match**: For each text chunk, we check if it can match to a concept in the ontology by various methods. Matched chunks will be saved as weak annotations for training our models.

Matching chief complaint mentions to corresponding concepts in an ontology is the core of this algorithm. To seek a balance between precision and recall of the weak labels, we employ three matching methods as a pipeline, as illustrated in Figure 1a: *S.1 - Exact string match*. We simply check if a chunk exactly matches to any alternative form of a concept in HaPPy ontology; *S.2 - Approximate string matching*. We use QuickUMLS (Soldaini and Goharian, 2016), a tool for extracting medical concepts using an approximate dictionary matching algorithm. This helps to resolve misspelling and lexical variations. *S.3 - Embedding-based matching*. It is achieved by computing the cosine similarity between the embedding of the chunk and that of an ontology concept. The embedding is obtained with fastText (Joulin et al., 2016), trained on our chief complaint corpus. In this way, a large amount

Tokens	Adjusted Token Weight	Adjusted Target Vector (B/I/O)
chest pain and	0.9	[0.9, 0.0, 0.1]
chest pain and	0.9	[0.0, 0.9, 0.1]
chest pain and	0.0	[0.0, 0.0, 1.0]

Table 3: An example of softened target labels for a chief complaint record consisting of three tokens. Both "chest" and "pain" can be matched to a concept in HaPPy ontology while "and" is a noisy token mistakenly matched.

of "weak" annotations of entity spans and concept labels are collected.

3.2 Entity Mention Extraction

We formulate the task of detecting entity mentions from a chief complaint record as a sequence labeling problem following the BIO tagging scheme, which is common in NER tasks. Each token takes a label, where "B/I" denotes a beginning/inside token of entity mentions, and "O" denotes other tokens outside mentions or separator tokens.

A pre-trained BERT (Devlin et al., 2019b) is used as the backbone of the model for contextualized language information, and a softmax layer is utilized on the top of BERT to classify the B/I/O tag of each token position. We also experimented with conditional random fields (CRF, Lafferty et al., 2001) on the top to learn the constrains among output tags. However, this extra CRF layer did not show advantageous performance so we dropped it.

One important feature of our model is the usage of the label smoothing technique to counter the noise in generated weak labels. With the help of label smoothing, the model can be more aware of the qualify of target labels so it can learn accordingly. Label smoothing (Szegedy et al., 2016) is a widely used technique to improve the generalization of neural network models (Müller et al., 2019).

We adjust the label smoothing to accommodate the weak span labels, since the similarity score in matching indicates the confidence level of a weak label. Formally, a target label can be represented as a three-dimensional vector where the value of each dimension follows $[p(B), p(I), p(O)], p(\cdot) \in [0, 1], \sum p(\cdot) = 1$. For each word in a chunk, we set the probability of a weak target label as the similarity between a chunk and its corresponding ontology concept. An example of label smoothing is shown in Table 3. We refer to our models for Mention Extraction in chief complaints as CCME.

3.3 Linking Entities to Ontology

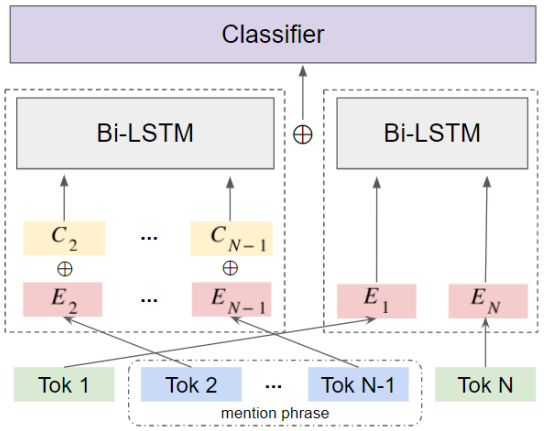


Figure 2: Architecture of the model for entity linking in chief complaints (CCEL) .

Each extracted entity mention will be linked with a given ontology through a classification model. We concatenate both word embedding and character embedding of an entity mention as the input, to accommodate the variable forms of medical entities. BERT embedding is not considered here as it's hard to be integrated with character embedding. Besides, two BiLSTMs are used to encode the tokens in its context from both directions, since the context before/after the mention may capture different information. Lastly, the outputs from both BiLSTMs are concatenated and fed to a softmax layer to obtain the concept label in the ontology. The diagram of our entity linking model called CCEL is shown in Figure 2.

4 Experimental Setup

4.1 Data

Our dataset[1] contains 1,232,899 free-text chief complaint records (in English) of patients' visits at 15 emergency departments of local healthcare institutes, including critical access hospitals, community hospitals, and tertiary care referral centers. All these departments use the same electronic health record system, but do not mandate a specific data entry format. The time span of the data is from 2015 to 2017. The dataset has been de-identified and approved by IRB.

HaPPy ontology is utilized as the target ontology to link entities. Additionally, we reduce the size of HaPPy ontology to 501 medical concepts

[1]All code/data for reproducing the results will be released at https://github.com/anonymous_repo, under IRB restrictions.

Models	Partial Match			Exact Match		
	Precision	Recall	F1	Precision	Recall	F1
†S&M (S.1)	95.81	36.90	53.28	92.69	35.70	51.54
†S&M (S.1 + S.2)	81.15	46.04	58.75	67.46	38.28	48.84
†S&M (S.1 + S.2 + S.3)	69.64	**57.36**	62.91	55.66	45.84	50.27
CCME-LSTM	78.45	48.26	59.76	67.25	42.39	52.00
CCME-BERT	81.37	53.43	64.50	71.46	48.52	57.80
CCME-BERT (soft)	83.41	56.70	**67.51**	72.95	**49.59**	59.04
CCME-ClinicalBERT (soft)	83.35	56.46	67.32	72.94	49.41	58.91
CCME-BERT (soft) + S&M (S.1)	**96.28**	44.86	61.20	**92.94**	43.30	**59.08**
CCME-BERT (soft) + S&M (S.1 + S.2)	86.13	51.82	64.71	76.86	46.24	57.74

Table 4: Entity extraction performance of different models. Scores are computed in **Partial Match** and **Exact Match** mode of SemEval'13. The **best**/2nd-best scores in each column are in bold/underlined. †S.1, S.2, S.3 refer to string matching algorithms in Figure 1(a), respectively.

(originally 692) by removing child nodes that have no significant clinical difference with their parent node. For example, two child nodes "ruq abdominal pain" ("ruq" means right upper quadrant) and "rlq abdominal pain" ("rlq" means right lower quadrant) are merged to their parent node "abd pain".

The testset consists of 1,013 instances (randomly sampled from the original collection) and 1,771 chief complaint mentions. Two domain experts independently annotated the data (linking mentions to concepts in HaPPy ontology), whereas the 2nd expert only annotated 200 data points to measure the inter-annotator agreement. The resulting Cohen's Kappa of concept classification (entity type) is 0.9326, and the accuracy of exact span overlap is 0.9029, both indicating a very high reliability between two annotators.

4.2 Experiment Settings

With regard to the entity mention extraction, we design three weak baselines based on our split-and-match (S&M) algorithm introduced in Sec 3.1: S&M (HaPPy), S&M (QuickUMLS) and S&M (Embedding). Besides, we also experiment with LSTM and several variants of BERT for CCME: CCME-LSTM, CCME-BERT and CCME-ClinicalBERT. We refer to the CCME-BERT model with the label smoothing as CCME-BERT (soft).

As for entity linking, a fastText- and a BERT-based classification model are used as baseline models. The input to fastText and BERT is the same as CCEL, i.e. mention tokens with context tokens within a window size of 2 (two tokens before/after the mention). We also train a fastText model following the single-label paradigm, that only predicts one concept class for each chief complaint record. We take MedType (Vashishth et al., 2020) (pretrained with WikiMed, PubMedDS and EHR documents, use QuickUMLS for extraction) as a strong baseline system.

We follow the setting of (Yang et al., 2018) for training the CCME-LSTM model and default settings of HuggingFace (Wolf et al., 2020) for BERT. Our CCEL model adopts two separate BiLSTMs, and the size of word-/character-level embedding is set to 100/30. The word-level embedding is initialized by training a fastText model on our CC corpus. All experiments are performed on a single NVIDIA 2080Ti graphics card (11GB).

4.3 Evaluation Metrics

We adopt the evaluation protocol of SemEval 2013 task 9.1 (Segura Bedmar et al., 2013) to evaluate models for mention extraction and entity linking. Precision, Recall, and F1 scores are reported in "Partial" mode (partial boundary match, regardless of the type) and "Exact" mode (exact boundary match, regardless of the type) for entity extraction. Besides, scores in "Entity type" mode (i.e., partial boundary match and correct entity type) are used for evaluating extraction and linking together.

5 Results and Discussion

5.1 Results of Entity Mention Extraction

Results of entity extraction are shown in table 4. Among three matching-based methods, S&M (HaPPy) has highest precision but lowest recall. This indicates that alternative strings provided by HaPPy ontology can lead to precise matches, but its coverage is very limited. Both S&M (QuickUMLS) and S&M (embedding) can achieve a better recall with the help of approximate string matching and embedding-based matching. But S&M (em-

#	Extraction Model	Linking Model	Precision	Recall	F1
s.1	†S&M (S.1)	†S&M (S.1)	**98.02**	37.75	54.51
s.2	†S&M (S.1 + S.2)	†S&M (S.1 + S.2)	79.34	45.02	57.44
s.3	†S&M (S.1 + S.2 + S.3)	†S&M (S.1 + S.2 + S.3)	57.73	47.54	52.14
n.1	CCME-BERT (soft)	fastText (single-label)	92.93	21.11	34.41
n.2	CCME-BERT (soft)	fastText	89.27	26.69	41.09
n.3	CCME-BERT (soft)	BERT	83.48	34.52	48.84
n.4	CCME-BERT (soft)	CCEL	84.36	45.22	58.88
b.1	QuickUMLS	MedType (EHR)	53.32	23.09	32.23
m.1	†S&M (S.1 + S.2)	fastText (single-label)	89.07	15.77	26.79
m.2	†S&M (S.1 + S.2)	fastText	86.49	19.52	31.85
m.3	CCME-BERT (soft)	†S&M (S.1)	97.86	45.60	62.20
m.4	CCME-BERT (soft)	†S&M (S.1 + S.2)	85.64	51.53	64.34
m.5	CCME-BERT (soft)	†S&M (S.1 + S.2) + CCEL	86.28	**55.43**	**67.49**

Table 5: Entity linking performance. Scores are computed in **Entity Type** mode of SemEval'13. The **best**/2nd-best scores in each column are in bold/underlined. †S.1, S.2, S.3 refer to string matching algorithms in Figure 1(a).

	P	R	F1
CCEL	**84.36**	**45.22**	**58.88**
- context emb	83.53	36.75	51.04
- character emb	83.46	29.35	43.43

Table 6: Ablation study on CCEL model with different input settings (using CCME-BERT(soft) as extraction model). Scores are computed following the **Entity type** setting of SemEval'13.

bedding) also introduces a fairly high rate of false positives and leads to the worst precision.

Most neural models outperform the matching-based methods, indicating that machine learning models can learn task-relevant inductive bias from weak labels. CCME-BERT performs better than LSTM, largely due to the better generalizability of BERT from pre-training. With the help of label smoothing, CCME-BERT (soft) demonstrates better performance than CCME-BERT model, suggesting that adjusting label weights by confidence can effectively alleviate the noise in weak labels. CCME-ClinicalBERT, which is pre-trained with electronic health record (EHR) notes, performs slightly worse than the CCME-BERT, suggesting the pre-training on EHR notes is not helpful for chief complaint related tasks. Without loss of generality, we mainly use CCME-BERT in following experiments.

We also consider the ensemble of CCME-BERT models and matching methods. Given the output (extracted mention spans) of CCME-BERT soft model, we further apply exact and approximate string matching aiming for a better precision. Compared with CCME-BERT (soft), both ensemble models achieve worse scores in terms of partial match, while CCME-BERT (soft) with S&M (HaPPy) performs the best for exact match. Although the overall performance is not improved, such ensemble models might useful when exact span matching is crucial to the task.

5.2 Results of Entity Linking

We present the scores of entity linking in Table 5, Among three matching-based methods, S&M (HaPPy) attains a high precision but a low recall. This confirms the observation from the extraction part. However, most neural models (n.1 to n.3) fail to beat the matching methods (s.1 to s.3) except for CCEL. Neural models can achieve a high precision score but the recall is relatively low. In the case of entity linking, weakly supervised labels cannot cover all the classes and most of them may concentrate on a small proportion of classes, thus models trained with weak labels may not perform well on classes that only have a small number of labels. Besides, different from entity extraction that models can utilize contextual information to detect spans, it is difficult for linking models to understand a target class if it is rarely seen in the data.

The major difference between CCEL and other neural models is on the direct use of contextual information. As shown in Table 6, the performance of CCEL drops after removing the context embedding part or the character embedding part, especially for recall. This confirms that including context information or character-level input can improve the model performance. As for fastText models, the single-label fastText model achieves highest precision among all neural models. Its advantage may lie in that the model is trained with the most likely labels and thus it is less affected by the noisy labels. Nonetheless, it fails to predict other concepts and results in the lowest recall.

Among all models, the general medical entity

Training	P	R	F1
WeakSup	**83.41**	56.70	67.51
Supervised	77.76	**89.66**	83.29
Fine-tuning	82.25	85.98	**84.07**

Table 7: Extraction performance (Partial Match) of CCME-BERT (soft) with three training strategies.

Train Data	Partial	Exact
w/ punct	25.98	9.91
w/o punct	34.17	17.63
w/ punct + denoising	**54.18**	**41.94**

Table 8: The extraction performance (F1) of CCME-BERT (soft) model on the *no-punctuation* test subset.

linking model MedType performs the worst, although it is pretrained with large medical datasets and annotated EHR documents. This indicates a non-negligible domain difference between chief complaint and common EHR data. Among five combined models, using CCME-BERT (soft) for extraction and S&M for linking (m.3, m.4) outperforms the reverse ways (m.1, m.2) by a clear margin. This confirms that CCME-BERT models are good at identifying entity mentions, while matching methods are good at classifying concepts given identified mentions. We also experiment with stacking CCEL to handle the unmatched cases of m.4 and it leads to a boost in recall.

5.3 Effect of Weak Supervision

In order to examine the effect of weak supervision, we conduct an ablation study by training models with and without weak labels. To this end, we conduct a five-fold cross validation and report the average score of five runs, by taking 80% of data points from the annotated test set as a training set for fully supervised learning, and the rest 20% is used for testing. We also experiment with two training strategies: 1) Supervised: training models with annotated data only; 2) Fine-tuning: pre-train the model using weak labels and fine-tune it with the annotated data. The result in Table 7 demonstrates that, even trained with little annotated data, CCME-BERT can achieve decent results on mention extraction, but pre-training the model with weak labels can be beneficial for precision and F1.

5.4 Resolving Cases without Punctuation Separators

We observe that in the chief complaint corpus nearly 40% of chief complaint records do not contain any punctuation as separators. The split-and-match algorithm cannot work to provide weak labeled data, and the models may fail to generalize to this kind of cases. To prove this hypothesis, we split the training data to two parts by checking if it contains punctuation separators. We train CCME-BERT with the two parts and evaluate models on a subset of the test set that only contains *no-punctuation* instances. The result in Table 8 shows that the model trained with labels derived from *with-punctuation* records does not generalize well to no-punctuation cases.

To improve models' robustness against this issue, we propose to train the model with additional noise: we intentionally remove punctuation marks from the instances that contain them. In this way, the model is trained to "denoise" each record by implicitly predicting the existence of separators. The resulting model demonstrates significant improvements over the previous ones.

5.5 Error Analysis

The majority of errors can be grouped into two categories. First, the absence of separators in a record leads to many wrong predictions. For example, in the record of "cough chest pain/congestion", the separator to delimit "cough" and "chest pain" is missing. Thus S&M methods are likely to miss one of the two chunks in mention extraction. Second, the variation in terminology can cause difficulty in linking. For example, we do not find the word "dysuria" in the dataset, since it's often referred to as "urinary tract pain". Even though neural models have the capability to infer synonyms, they can hardly perform well without a certain amount of annotated data to learn such knowledge.

6 Conclusions

We propose WESEEL, a weak supervision method for extracting and linking chief complaint entities in the absence of human annotation. We develop a split-and-match algorithm to produce weak labels of both entity spans and concept labels. We also show that the customized label smoothing can effectively alleviate the noise in weak labels. Our framework is considered to be generic enough for chief complaint data across departments and we will check its generalizability on more data in the future. Our framework is also applicable to other entity identification tasks where human annotation is limited, and we leave it for future work.

References

Dominik Aronsky, Diane Kendall, Kathleen Merkley, Brent C James, and Peter J Haug. 2001. A comprehensive set of coded chief complaints for the emergency department. *Academic emergency medicine*, 8(10):980–989.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Parminder Bhatia, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. 2019. Comprehend medical: a named entity recognition and relationship extraction web service. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1844–1851. IEEE.

Lidong Bing, Sneha Chaudhari, Richard C Wang, and William Cohen. 2015. Improving distant supervision for information extraction using label propagation through lists. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 524–529.

David Chang, Woo Suk Hong, and Richard Andrew Taylor. 2020. Generating contextual embeddings for emergency department chief complaints. *JAMIA open*, 3(2):160–166.

Lihu Chen, Gaël Varoquaux, and Fabian M. Suchanek. 2021. A lightweight neural model for biomedical entity linking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12657–12665.

Mike Conway, John N Dowling, and Wendy W Chapman. 2013. Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in north america. *Journal of biomedical informatics*, 46(4):734–743.

Jagan Dara, John N. Dowling, Debbie Travers, Gregory F. Cooper, and Wendy W. Chapman. 2008. Evaluation of preprocessing techniques for chief complaint classification. *Journal of Biomedical Informatics*, 41(4):613–623.

Mostafa Dehghani, Arash Mehrjou, Stephan Gouws, Jaap Kamps, and Bernhard Schölkopf. 2017a. Fidelity-weighted learning. *CoRR*, abs/1711.02799.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017b. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Jarunee Duangsuwan and Pawin Saeku. 2018. Semiautomatic classification based on icd code for thai text-based chief complaint by machine learning techniques. *International Journal of Future Computer and Communication*, 7(2).

Marta Fernandes, Rúben Mendes, Susana M Vieira, Francisca Leite, Carlos Palos, Alistair Johnson, Stan Finkelstein, Steven Horng, and Leo Anthony Celi. 2020. Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing. *PloS one*, 15(3):e0229331.

Nathaniel R Greenbaum, Yacine Jernite, Yoni Halpern, Shelley Calder, Larry A Nathanson, David A Sontag, and Steven Horng. 2019. Improving documentation of presenting problems in the emergency department using a domain-specific ontology and machine learning-driven user interfaces. *International journal of medical informatics*, 132:103981.

Stephanie W Haas, Debbie Travers, Judith E Tintinalli, Daniel Pollock, Anna Waller, Edward Barthell, Catharine Burt, Wendy Chapman, Kevin Coonan, Donald Kamens, et al. 2008. Toward vocabulary control for chief complaint. *Academic Emergency Medicine*, 15(5):476–482.

Michael A Hedderich and Dietrich Klakow. 2018. Training a neural network in a low-resource setting on automatically annotated noisy data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 12–18.

Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. 2018. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468*.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 541–550.

Steven Horng, Nathaniel R Greenbaum, Larry A Nathanson, James C McClay, Foster R Goss, and

Jeffrey A Nielson. 2019. Consensus development of a modern ontology of emergency department presenting problems—the hierarchical presenting problem ontology (happy). *Applied clinical informatics*, 10(03):409–420.

Jia-Hao Hsu, Ting-Chia Weng, Chung-Hsien Wu, and Tzong-Shiann Ho. 2020. Natural language processing methods for detection of influenza-like illness from chief complaints. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1626–1630. IEEE.

William Hsu, Simon X Han, Corey W Arnold, Alex AT Bui, and Dieter R Enzmann. 2016. A data-driven approach for quality assessment of radiologic interpretations. *Journal of the American Medical Informatics Association*, 23(e1):e152–e156.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Sotiris Karagounis, Indra N. Sarkar, and Elizabeth S. Chen. 2020. Coding free-text chief complaints from a health information exchange: A preliminary study. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Scott H Lee, Drew Levin, Patrick D Finley, and Charles M Heilig. 2019. Chief complaint classification with recurrent neural networks. *Journal of biomedical informatics*, 93:103158.

Dingcheng Li, Sijia Liu, Majid Rastegar-Mojarad, Yanshan Wang, Vipin Chaudhary, Terry Therneau, and Hongfang Liu. 2016. A topic-modeling based framework for drug-drug interaction classification from biomedical text. In *AMIA Annual Symposium Proceedings*, volume 2016, page 789. American Medical Informatics Association.

Katherine P Liao, Tianxi Cai, Guergana K Savova, Shawn N Murphy, Elizabeth W Karlson, Ashwin N Ananthakrishnan, Vivian S Gainer, Stanley Y Shaw, Zongqi Xia, Peter Szolovits, et al. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *bmj*, 350.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, pages 4694–4703. Curran Associates, Inc.

Aurélie Névéol, Pierre Zweigenbaum, et al. 2017. Making sense of big textual data for health care: findings from the section on clinical natural language processing. *Yearbook of medical informatics*, 26(1):228.

John D. Osborne, James S. Booth, Tobias O'Leary, Amy Mudano, Giovanna Rosas, Phillip Foster, Kenneth Saag, and Maria Danila. 2020. Identification of gout flares in chief complaint text using natural language processing. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

C Rochefort, A Verma, T Eguale, and D Buckeridge. 2015. O-037: surveillance of adverse events in elderly patients: a study on the accuracy of applying natural language processing techniques to electronic health record data. *European Geriatric Medicine*, (6):S15.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 959–962.

Feichen Shen, Sijia Liu, Yanshan Wang, Liwei Wang, Naveed Afzal, and Hongfang Liu. 2017. Leveraging collaborative filtering to accelerate rare disease diagnosis. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1554. American Medical Informatics Association.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Fei Teng, Zheng Ma, Jie Chen, Ming Xiao, and Lufei Huang. 2020. Automatic medical code assignment via deep learning approach for intelligent healthcare. *IEEE journal of biomedical and health informatics*, 24(9):2506–2515.

Debbie Travers, Shiying Wu, Matthew Scholer, Matt Westlake, Anna Waller, and Anne-Lyne McCalla. 2007. Evaluation of a chief complaint pre-processor for biosurveillance. In *AMIA annual symposium proceedings*, volume 2007, page 736. American Medical Informatics Association.

Debbie A Travers and Stephanie W Haas. 2006. Unified medical language system coverage of emergency-medicine chief complaints. *Academic emergency medicine*, 13(12):1319–1323.

Ilya Valmianski, Caleb Goodwin, Ian M Finn, Naqi Khan, and Daniel S Zisook. 2019. Evaluating robustness of language models for chief complaint extraction from patient-generated text. *arXiv preprint arXiv:1911.06915*.

Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, and Carolyn Rose. 2020. Medtype: Improving medical entity linking with semantic type prediction. *arXiv preprint arXiv:2005.00460*.

Michael M Wagner, William R Hogan, Wendy W Chapman, and Per H Gesteland. 2006. Chief complaints and icd codes. *Handbook of biosurveillance*, page 333.

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, 17(1):4572–4596.

Yipei Wang, Xingyu Fan, Luoxin Chen, I Eric, Chao Chang, Sophia Ananiadou, Junichi Tsujii, and Yan Xu. 2019. Mapping anatomical related entities to human body parts based on wikipedia in discharge summaries. *BMC bioinformatics*, 20(1):1–11.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. *arXiv preprint arXiv:2010.02405*.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 817–824.