Effective Feature Representation for Clinical Text Concept Extraction

Yifeng Tao
Roam Analytics
Carnegie Mellon University

Bruno Godefroy
Roam Analytics

Guillaume Genthial
Roam Analytics

Christopher Potts
Roam Analytics
Stanford University

Abstract

Crucial information about the practice of healthcare is recorded only in free-form text, which creates an enormous opportunity for high-impact NLP. However, annotated healthcare datasets tend to be small and expensive to obtain, which raises the question of how to make maximally efficient uses of the available data. To this end, we develop an LSTM-CRF model for combining unsupervised word representations and hand-built feature representations derived from publicly available healthcare ontologies. We show that this combined model yields superior performance on five datasets of diverse kinds of healthcare text (clinical, social, scientific, commercial). Each involves the labeling of complex, multi-word spans that pick out different healthcare concepts. We also introduce a new labeled dataset for identifying the treatment relations between drugs and diseases.

1 Introduction

The healthcare system generates enormous quantities of data, but its tools for analytics and decision-making rely overwhelmingly on a narrow subset of structured fields, especially billing codes for procedures, diagnoses, and tests. The textual fields in medical records are generally under-utilized or completely ignored. However, these clinical texts are our only consistent source of information on a wide variety of crucial factors – hypotheses considered and rejected, treatment rationales, obstacles to care, brand recognition, descriptions of uncertainty, social and lifestyle factors, and so forth. Such information is essential to gaining an accurate picture of the healthcare system and the experiences of individual patients, creating an enormous opportunity for high-impact NLP.

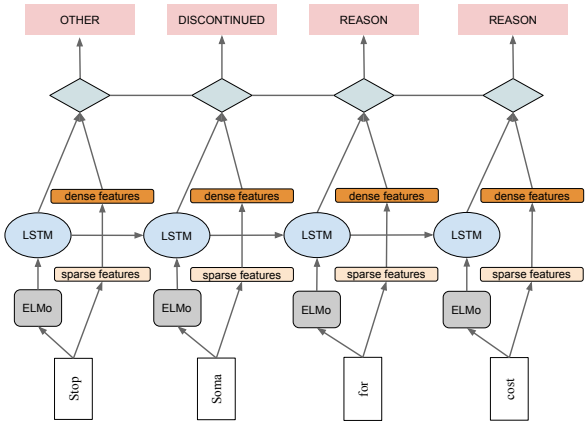

Figure 1: Model diagram. In our full model, words are represented by pretrained ELMo embeddings, which feed into LSTM cells, and by sparse ontology-derived feature representations, which are fed to a dense layer with dropout to produce a lower-dimensional representation that is concatenated with the hidden states of the LSTM. The resulting mixed feature representation is fed into a CRF layer that forms the basis for token-level label predictions. We assess this full model against variants without the LSTM or hand-built features to motivate the full version.

However, annotated clinical text datasets are scarce and tend to be small, for two reasons. First, data access is usually highly limited because of privacy considerations; the inherent richness of language data means that de-identification is hard or impossible (Uzuner et al., 2007). Second, because healthcare concepts are complex, the needed annotations generally must be provided by domain

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 1–14
Minneapolis, Minnesota, June 7, 2019.

specialists who are trained both in the practice of healthcare and in the interpretation of healthcare records. Such experts are in high demand, and the annotation work they do is intellectually challenging, so the annotated datasets they produce are, by any measure, very expensive. The result is that even the largest annotated clinical text datasets are small by comparison with those from other areas of NLP, and this has profound consequences for the kinds of models that are viable in this space.

In this paper, we define a hybrid LSTM-CRF model that is effective for real-world clinical text datasets. The architecture is sketched in figure 1. Its crucial property is that it synthesizes two kinds of feature representation: dense representations that can be trained on any large text corpus (not necessarily using clinical text) and sparse, high-dimensional feature representations based on hand-built feature functions. Hand-built feature functions are especially powerful in healthcare because they can leverage the numerous high-quality medical lexicons and ontologies that are publicly available. As a result, such features can achieve impressive coverage with relatively little additional effort.

We show that this combined model yields superior performance on five datasets of diverse kinds of healthcare text: two clinical, one social media, one scientific, and one commercial/regulatory (official drug labels). Each task involves the labeling of complex, multi-word spans that pick out diverse healthcare concepts: the Chemical–Disease Relation dataset (CDR; Wei et al. 2015); the Penn Adverse Drug Reaction Twitter dataset (ADR; Nikfarjam et al. 2015); a new disease diagnosis dataset; a new prescription reasons dataset that involves identifying complex REASON spans for drug–prescription actions; and a new dataset of 10K drug–disease treatment descriptions, which we release with this paper.

2 Models

Our full model is depicted schematically in figure 1. Its modular structure defines a number of variations that allow us to quantify the value of including dense and sparse feature representations obtained from diverse sources.

Individual words are represented in two ways in the full model: with dense, pretrained vectors and with sparse, high-dimensional feature representations derived from hand-built feature functions. If the dense representations are removed, the LSTM cells are also removed, resulting in a standard CRF (Lafferty et al., 2001; Sutton and McCallum, 2011). If the sparse representations are removed, the result is a standard LSTM-based RNN (Hochreiter and Schmidhuber, 1997).

We explore two ways of initializing the dense representations: random initialization according to the method of Glorot and Bengio (2010) and the ELMo embeddings released by Peters et al. (2018). The ELMo embeddings were trained on the 1 billion word benchmark of Chelba et al. (2013) – general newswire text not specialized to the healthcare space. What is special about ELMo embeddings, as compared to more standard word representation learning, is that they are obtained from the parameters of a full language model, so that each word's representation varies by, and is sensitive to, its linguistic context; see also McCann et al. 2017; Radford et al. 2018.

The nature of the hand-built feature representations varies by task, so we leave most of the details to section 3. All the models featurize each word in part using the word and part-of-speech tag of the current word and the preceding and following four words. They also include features that seek to characterize the nature of the semantic environment: markers of negation, uncertainty, hedging, and other core task-specific contextual cues. Finally, the feature functions make extensive use of drug and disease lexicons to identify the types of words. The drug lexicons are RxNorm, the National Drug Code (NDC), FDA Drug Labels, FDA Orange Book, and the OpenFDA fields found in a number of public FDA datasets (e.g., Drug Adverse Events). The disease lexicons are derived from historical ICD-9 and ICD-10 code sets, SNOMED-CT (Spackman et al., 1997), the Disease Ontology (Schriml et al., 2011; Kibbe et al., 2014), and the Wikidata graph (Vrandečić and Krötzsch, 2014). The wealth and diversity of these sources is typical of healthcare and highlights the potential for taking advantage of such resources to help overcome the challenges of small datasets. Table A1 shows an example of hand-built features.

In the full model, we include a dense layer that transforms the sparse feature representations, and we apply dropout (Hinton et al., 2012) to this layer. These transformed representations are concatenated with the hidden states of the LSTM to produce the full representations for each word.

Dataset	Example
Diagnosis Detection	Asymptomatic/POSITIVE bacteriuria/POSITIVE , could be neurogenic/CONCERN bladder/CONCERN disorder/CONCERN .
Prescription Reasons	I will go ahead and place him on Clarinex/PRESCRIBED for/REASON his/REASON seasonal/REASON allergic/REASON rhinitis/REASON .
Penn Adverse Drug Reactions (ADR)	#TwoThingsThatDontMixWell venlafaxine and alcohol- you'll cry/ADR and throw/ADR chairs/ADR at your mom's BBQ.
Chemical–Disease Relations (CDR)	Ocular/DISEASE and/DISEASE auditory/DISEASE toxicity/DISEASE in hemodialyzed patients receiving desferrioxamine/DRUG .
Drug–Disease Relations	Indicated for the management of active/TREATS rheumatoid/TREATS arthritis/TREATS and should not be used for rheumatoid/CONTRA arthritis/CONTRA in/CONTRA pregnant/CONTRA women/CONTRA .

Table 1: Short illustrative examples from each of our five datasets, with some modifications for reasons of space. CDR examples are typically much longer, encompassing an entire scientific title and abstract. Section 3 more fully explicates the labels. All unlabeled tokens are labeled with OTHER.

Where the hand-built representations are left out, the word representations are simply the hidden states of the RNN; where the dense representations are left out, the word representations are simply the sparse representations, resulting in a standard linear-chain CRF.

There is a natural variant of the model depicted in figure 1 in which the CRF layer is replaced by a softmax layer. In our experiments, this was always strictly worse than the CRF layer. Another variant feeds the compressed hand-built features together with ELMo embeddings into the LSTM. This too led to inferior or comparable performance. Finally, we evaluated a version that used a bidirectional LSTM, but found that it did not yield improvements. Therefore, we do not include those experimental results, to simplify the discussion.

3 Experiments

We report experiments on five different datasets: two from transcribed clinical narratives, one from social media, one from scientific publications, and one from official FDA Drug Labels texts. For each, the task is to label spans of text that identify particular healthcare concepts. We are particularly interested in the capacity of our models to identify multi-word expressions in a way that is sensitive to the semantics of the environment – for example, to distinguish between a drug prescribed and a drug discontinued, or to distinguish disease mentions as diagnoses, diagnostic concerns, or ruled-out diagnoses. Table 1 gives a short illustrative example from each dataset. Table A2 gives detailed statistics for each dataset.

Three of the datasets are already partitioned into training and test sets. For these, we tune the hyperparameters using 5-fold cross-validation on the training set, train the model with tuned hyperparameters on the training set, and then evaluate the performance of the trained model on the test set.

The other two datasets do not have predefined splits. For these, we divide them equally into five parts. For each fold, the hyperparameters are tuned on the training data (also using 5-fold cross-validation), and the best model is then applied to the test data for the evaluation. These experiments are repeated three times to smooth out variation deriving from the random initialization of the model parameters, though we use the hyperparameters selected for each fold in the first run in the subsequent two experiments to save computational resources.

We use the Adam optimizer (Kingma and Ba, 2014), with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, the training batch size set to 16, and the dropout rate set to 0.5 for all the experiments. The step size η and the coefficients of the ℓ_1 and ℓ_2 regularizers c_1 and c_2 are tuned. The step size is first tuned by setting both $c_1 = c_2 = 0$, and then c_1 and c_2 are tuned using random search (Bergstra and Bengio, 2012) for ten settings. Table A3 provides additional details on our hyperparameters and evaluation protocol.

The source code for our experiments and models is available.[1]

3.1 Diagnosis Detection

Our Diagnosis Detection dataset is drawn from a larger collection clinical narratives – de-identified transcriptions of the reports healthcare professionals record about their interactions with patients. The corpus was provided to us by a healthcare start-up. We sampled and labeled 6,042 sentences for information about disease diagnoses. The labels are POSITIVE DIAGNOSIS, CONCERN, RULED-OUT, and OTHER. The labeling was done by a team of domain experts. The challenging aspects of this task are capturing the complex, multi-word disease names and distinguishing the semantic sense of those mentions (as summarized by our label set) based on their sentential context.

For the hand-built parts of our representations, we extend the basic feature set described in section 2 with cue words that help identify whether a description is about a patient's history or current condition, as well as cue words for causal language, measurements, and dates. The power these features bring to the model, beyond what is captured in the ELMo-LSTM representations, is evident in table 2, column 1.

3.2 Prescription Reasons

Our Prescription Reasons dataset is drawn from the same corpus of clinical narratives as our Disease Diagnosis dataset and was annotated by the same team of domain experts. This dataset contains 5,179 sentences, with labels PRESCRIBED, DISCONTINUED, REASON, and OTHER. For the first two labels, the majority are unigrams naming drugs. Of special interest is the REASON category, which captures long, highly diverse reasons for actions taken concerning prescription drugs. (The relations are captured with additional edge annotations connecting spans, but we do not model them in this paper.) This information about the rationale for prescription decisions is the sort of thing that appears only in text, and it has clear value when it comes to understanding these decisions, making this an especially interesting task.

Our hand-built feature representations are similar to those used for Diagnosis Detection, but they additionally contain features based in large drug lexicons, as discussed in section 2, as well as features based on cue-words for different prescription actions: switching, discontinuing, increasing, decreasing, and so forth. The results in table 2, column 2, clearly favor the combined model that uses both these features and the ELMo-LSTM.

3.3 Penn Adverse Drug Reactions (ADR)

The Penn Adverse Drug Reactions (ADR; Nikfarjam et al. 2015) dataset is an annotated collection of tweets giving informal adverse reactions to prescription drugs. It's thus a different kind of clinical text than in our two previous experiments – public self-reports by patients, rather than private technical descriptions by healthcare professionals.

The original dataset contained 1,340 labeled tweets for training and 444 for testing. However, due to restrictions on redistributing Twitter data, the project team was unable to release the tweets, but rather only a script for downloading them. Due to tweet deletions, we were able to download only 749 train examples and 272 test examples. This limits our ability to compare against prior work on this dataset, but the small size further tests our hypothesis that our combined model can get traction with relatively few examples.

For our hand-built feature functions, we follow the protocol specified in the ADRMine CRF package released by Nikfarjam et al. (2015). Key components include tokenization (Gimpel et al., 2011), spelling correction (Cutting, 1999; Atkinson, 2018), lemmatization, and featurization (Loper and Bird, 2002). Thus our combined model is a strict extension of this publicly available package (setting aside differences related to implementation and optimization). We follow Nikfarjam et al. (2015) in using Inside/Outside/Beginning (IOB; Ramshaw and Marcus 1995) tags.

Our test-set results, given in table 2, column 3, show the power of our combined model. For context, the best results reported by Nikfarjam et al. are 72.1, for a CRF that includes hand-built features as well as features based on the cluster indices of distributional word representations. That is, their model draws on similar insights to our own. Though we only have half of the training samples, our unified model is still able to get traction on this dataset.

3.4 Chemical–Disease Relations (CDR)

The Biocreative V Chemical Disease Relation dataset of Wei et al. (2015) captures relationships

[1] https://github.com/roamanalytics/roamresearch/tree/master/Papers/Feature4Healthcare

	Diagnosis Detection	Prescription Reasons	Penn Adverse Drug Reactions (ADR)	Chemical–Disease Relations (CDR)	Drug–Disease Relations
rand-LSTM-CRF	77.3 ± 0.05	69.6 ± 0.25	53.8 ± 0.88	85.1 ± 0.10	48.2 ± 1.12
HB-CRF	82.0 ± 0.05	78.5 ± 0.01	58.8 ± 0.12	86.2 ± 0.02	42.3 ± 0.30
ELMo-LSTM-CRF	83.9 ± 0.35	81.0 ± 0.20	65.7 ± 0.35	88.2 ± 0.34	50.6 ± 0.64
ELMo-LSTM-CRF-HB	**85.3** $\pm 0.24^{***}$	**82.0** $\pm 0.03^{***}$	**68.5** $\pm 1.67^{*}$	**89.9** $\pm 0.12^{***}$	**51.9** $\pm 0.52^{**}$

Table 2: Per-token macro-F1 scores. For ADR, the F1 scores are for chunks via approximate matching (Nikfarjam et al., 2015; Tsai et al., 2006). 'rand-LSTM' is an LSTM with randomly initialized word vectors. 'ELMo-LSTM' is an LSTM initialized with pretrained ELMo embeddings. 'HB' signals sparse, high-dimensional feature representations based on hand-built feature functions. The mean values and standard deviations are calculated using F1 scores of three runs of repeated experiments, as discussed in section 3. Statistical significance notation for the last two rows (two top-performing models) is *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

between chemicals and diseases in the titles and abstracts for scientific publications. It contains 1,000 training texts and 500 test texts. Its labels are CHEMICAL, DISEASE, and OTHER. This dataset is not only from a different domain than our others, but it also involves much longer texts.

Our hand-built feature function is exactly the one used for the Prescription Reasons experiments. We report results for the standard test set. The power of the combined model is again evident in the results in table 2, column 4.

3.5 Drug–Disease Relations

Our final experiments are on a new annotated dataset that we will be releasing along with this paper.[2] The underlying corpus is FDA Drug Labels, which contains all the official labels for all drugs licensed for sale in the U.S. These labels include a wide range of information, including active ingredients, warnings, and approved usages. Our annotation project focused on capturing the relationship between these drugs and mentioned diseases. The resulting labels are TREATS, PREVENTS, UNRELATED and CONTRAINDICATED-FOR. Figure A1 describes the corpus-building process in more detail.

Since FDA Drug Labels is a public dataset, we used this as an opportunity to see whether we could obtain good labels via crowdsourcing. This effort proceeded in two phases. In the first, annotators identified disease spans, working from an annotation manual that provided guidance on how to delimit such phrases and lexical resources to help them identify diseases. In the second phase, annotators assigned the span labels from our label set, again using an annotation manual we created to guide their choices.

We launched our task on Figure Eight with 10,000 sentences. It was completed within a few days. The job was done by 1,771 people from 72 countries, the majority from Venezuela. No special qualifications were imposed. To infer a label for each example, we applied Expectation Maximization (EM), essentially as in Dawid and Skene (1979). The inter-annotator agreement between these labels and those we inferred via EM is 0.83 for both tasks. For assessment, a team of experts independently labeled 500 examples from the same pool of sentences, using the same criteria and annotation manuals as the crowdworkers. The inter-annotator agreement between the labels inferred from the crowd and those from the experts is 0.82, suggesting that the inferred labels are good.

We expect the crowdsourced labels to be used only for training. Our test set consists entirely of non-train examples with labels assigned by experts. This allows us to train on noisy labels, to check for robustness, while still assessing on truly gold labels. Our results for this experiment are given in table 2, column 5, and point to the superiority of our combined model.

4 Discussion

Our discussion seeks to show that the combined model, which shows superior performance in all tasks (table 2), is making meaningful use of both kinds of features (hand-built and ELMo) and both of the major model components (LSTM and CRF).

4.1 The Role of Text Length

We expect the LSTM to handle short texts very effectively, but that its performance will be degraded for long ones. In contrast, the CRF might fall short of the LSTM on short texts, but it should be more

[2] https://github.com/roamanalytics/roamresearch/tree/master/BlogPosts/Features_for_healthcare

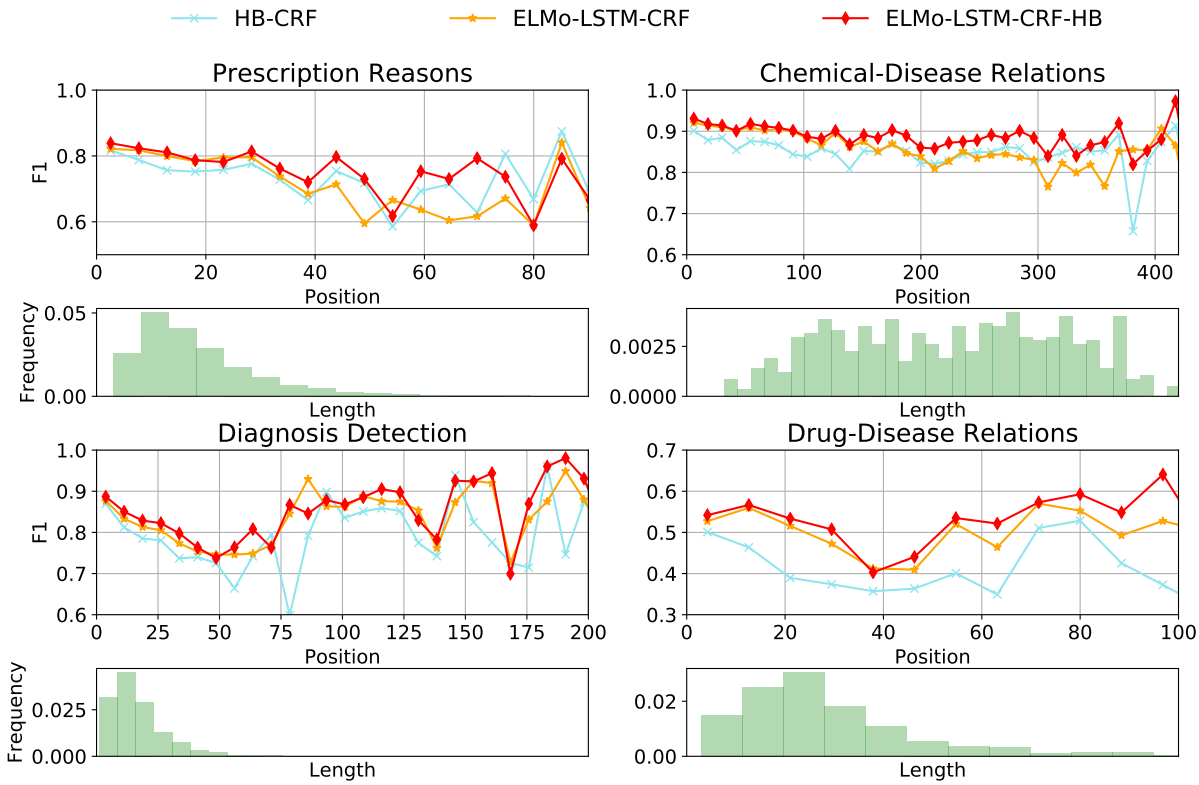

Figure 2: Text-length experiments. Along with the distribution of text lengths, per-token macro-F1 scores of words that fall into specific bins in the sentences are shown. For the top two datasets, the ELMo-LSTM-CRF is better at earlier positions, while the HB-CRF is better at later ones. For the bottom two datasets, the ELMo-LSTM-CRF is always better than the HB-CRF. In all these cases, the combined model takes advantage of both models and always outperforms the base models. ADR dataset results are given in figure A2 due to space limitations.

robust on long ones. We thus hypothesize that the combined model will learn to take advantage of these comparative strengths.

We find strong support for this hypothesis in our data. Figure 2 illustrates this. These plots track the macro-F1 scores (y-axes) of tokens in specific linear positions (x-axes). There are two major trends.

First, in the Prescription Reasons and CDR datasets (top two panels), we see that the HB-CRF starts to outperform the ELMo-LSTM-CRF after about word 40 in Prescription Reasons (which contains many long texts that list patient history; section 3.3) and after about word 160 in CDR (which has paragraph-length texts; section 3.4).

Second, in the Diagnosis Detection and Drug–Disease Relations datasets (bottom two panels in figure 2), the ELMo-LSTM-CRF model outperforms the HB-CRF at all positions. However, there is still evidence that our full model is leveraging the strengths of both of its major components, as it outperforms both in all positions.

In summary, the performance curve of the combined model is roughly an upper envelope of the two base-model curves. The combined model is able to achieve better performance for both short and long texts, and for words in any position, by utilizing features from both base models.

4.2 Analysis of the CRF Potential Scores

The potential scores (also referred to as "unary scores" or "emissions" in some work) of the CRF provide another method for model introspection. These scores are the direct inputs to the final CRF layer, where the token-level label predictions are determined. When the potential score for a specific label is high, the CRF assigns a high weight to that label under the contraints of adjacent labels. Thus, by checking the potential scores for the feature dimensions deriving from each of our base models, we can gain insights into the relative importance of these models and how the combined model leverages features from both.

The potential scores of each word in the test set are shown in figure 3, where the left panels show the LSTM features and the right panels show the CRF (hand-built) features. Due to the general ef-

fectiveness of the ELMo-LSTM, we always have higher average potential scores from those features. This is reflected in the mean scores at left and in the comparatively large amount of white (high scores) in the panels. However, the hand-built features always make substantial contributions, especially in Diagnosis Detection, Prescription Reasons, and CDR. We note also that, where the performance of the two base models is very similar (table 2), the potential scores in the combined model are also more similar.

4.3 Major Improvements in Minor Categories

One of our central motivations for this work is that clinical datasets tend to be small due to the challenges of getting quality labels on quality data. These size limitations impact model performance, and the hardest hit categories tend to be the smallest ones. Unfortunately, these are often the most important categories, identifying rare but significant events. We are thus especially interested in whether our combined model can address this problem.

Table 3 suggests that the combined model does make progress here, in that the largest gains, across all relevant datasets, tend to be for the smallest categories. This is very dramatically true for the Drug–Disease Relations dataset, where only the combined model is able to get any traction on the smallest categories; it achieves 103.5% and 71.3% improvements in F1 score over the HB-CRF model for the two smallest categories. It seems clear that, in transferring compact embedding representations learned from other large text datasets, the combined model can elevate performance on small categories to an acceptable level.

5 Prior Work

5.1 Clinical Text Labeling

Apache cTAKEs (Savova et al., 2010) extracts information from clinical text. Its labeling module implements a dictionary look-up of concepts in the UMLS database, and the concept is then mapped into different semantic types (labels). Similar extractions play a role in our hand-built features, but only as signals that our models learn to weight against each other to make decisions.

ADRMine (Nikfarjam et al., 2015) is closer to our own approach; it focuses on extracting adverse drug reaction mentions from noisy tweets. It combines hand-built features and word embedding cluster features for label prediction. However, our model is more powerful in the sense that we directly utilize the word embeddings and feed them into the LSTM.

Habibi et al. (2017) use a combined LSTM-CRF to achieve better NER results on 33 biomedical datasets than both available NER tools and entity-agnostic CRF methods, though they do not incorporate hand-built features.

There are also competitions related to labeling tasks in the context of clinical text. The i2b2 Challenge (Sun et al., 2013) includes event detection as one of the task tracks, which is basically a labeling task. The best results on this task came from a team using a simple CRF. The Biocreative V Chemical–Disease relation (CDR) competition (Wei et al., 2015) released a widely used dataset for researchers to evaluate their NER tools for biomedical text, and Verga et al. (2018) report state-of-the-art results for a self-attention encoder, using a dataset that extends CDR.

5.2 Efficient Annotation

Obtaining accurate annotations is expensive and time consuming in many domains, and a rich line of research seeks to ease this annotation burden. Ratner et al. (2016) and Hancock et al. (2018) propose to synthesize noisy labeling functions to infer gold training labels, and thus make better use of annotators' time, by allowing them to focus on writing high-level feature functions (and perhaps label individual examples only for evaluation). These efforts are potentially complementary to our own, and our experiments on our new Drug–Disease dataset (section 3.5) suggest that our combined model is especially robust to learning from noisy labels compared with base models.

5.3 Related Models

A large body of work explores combined LSTM and CRF models for text labeling. Huang et al. (2015) use an LSTM-CRF for sequence tagging, and Ma and Hovy (2016) propose a bi-directional LSTM-CNNs-CRF for the same task. In addition to word embeddings, Lample et al. (2016) utilize character embedding information as the input to a LSTM-CRF. Jagannatha and Yu (2016) integrate pairwise potentials into the LSTM-CRF model, which improves sequence-labeling performance in clinical text. Wang et al. (2018) and Crichton et al. (2017) use multi-task learning based on the ba-

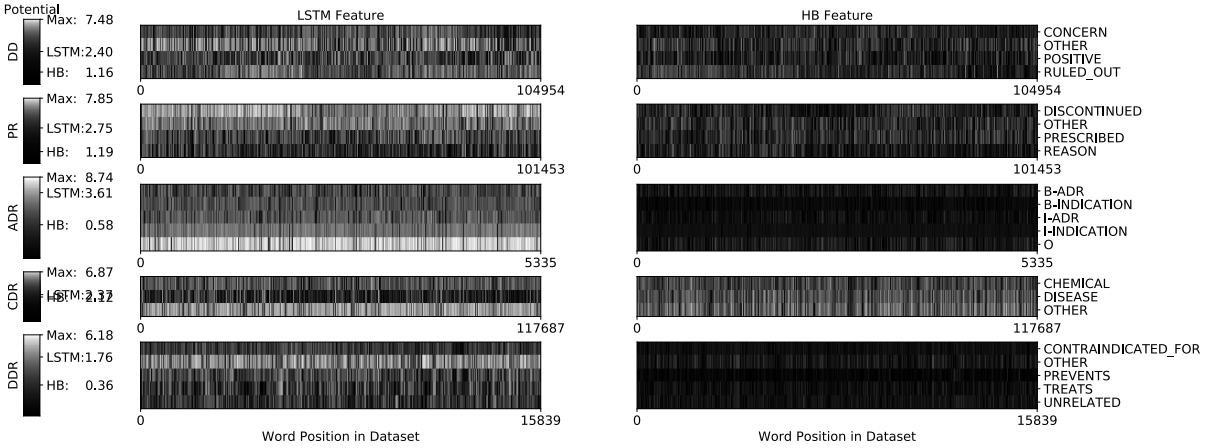

Figure 3: Potential score experiments. Potential scores from the ELMo-LSTM and HB modules of all five datasets are shown. Mean potential scores of both features are calibrated in the left colorbar. Higher potential scores (lighter cells) indicate greater importance for the feature. In all five datasets, the combined model pays more attention to the ELMo-LSTM features, but the hand-built features always contribute. Comparing with the results in table 2, we note that when the performance of two base models is comparable, their potential scores in the combined model are also closer.

Diagnosis Detection				Prescription Reasons			
Label	Support	F1 score	Improvement	Label	Support	F1 score	Improvement
OTHER	74888	95.3	1.4%	OTHER	83618	95.8	0.9%
POSITIVE	24489	86.1	4.4%	REASON	9114	64.7	8.6%
RULED-OUT	2797	86.4	3.6%	PRESCRIBED	5967	84.7	4.4%
CONCERN	2780	72.1	5.6%	DISCONTINUED	2754	82.7	5.6%
Chemical–Disease Relations (CDR)				**Drug–Disease Relations**			
Label	Support	F1 score	Improvement	Label	Support	F1 score	Improvement
OTHER	104530	98.3	0.5%	OTHER	10634	90.8	2.3%
DISEASE	6887	84.2	6.3%	TREATS	3671	76.0	5.7%
CHEMICAL	6270	87.0	6.7%	UNRELATED	1145	53.8	71.3%
				PREVENTS	320	41.1	103.5%
				CONTRAINDICATED-FOR	69	0	–

Table 3: Relative F1 score improvements of different labels. For each label, we give the number of supporting examples (Support), the F1 score of our combined model, and the relative improvements over the HB-CRF model. The F1 scores of minor labels suffer from insufficient training data, and thus have lower values. However, the combined model shows the largest relative improvements in these categories. ADR results are shown in table A4.

sic LSTM-CRF structure to improve NER performance in biomedical text. Our model provides an effective method for fully utilizing the sparse ontology-driven features left out of by the above work, which are complementary to dense embeddings and therefore boost performance of clinical concept extraction with limited training data (section 4).

There are also a number of models that mix dense and sparse feature representations. Gormley et al. (2015) and Cheng et al. (2016) combine both unlexicalized hand-crafted features and word embeddings to improve the performance of relation extraction in recommender systems. However, they focus on simple multi-layer perceptron models, rather than considering a more expressive LSTM structure. Similarly, Wang et al. (2019) utilize both sparse UMLS features and unpretrained word embeddings as the input to an LSTM for genetic association inferences from medical literature. While their UMLS features are a single look-up table of semantic types, our model relies on much richer resources of medical knowledge and includes more heterogeneous and expressive hand-built features that capture the semantic, morphological and contextual information of words (section 2).

6 Conclusion

Clinical text datasets are expensive to label and thus tend to be small, but the questions they can answer are often very high-impact. It is thus incumbent upon us to make maximally efficient use of these resources. One way to do this is to draw heavily on lexicons and other structured resources to write feature functions. Another way is to leverage unlabeled data to create dense feature vectors.

The guiding hypothesis of this paper is that the best models will make use of both kinds of information. To explore this hypothesis, we defined a new LSTM-CRF architecture that brings together these two kinds of feature, and we showed that this combined model yields superior performance on five very different healthcare-related tasks. We also used a variety of introspection techniques to gain an understanding of how the combined model balances its different sources of information. These analyses show that the combined model learns to pay attention to the most reliable sources of information for particular contexts, and that it is most effective, as compared to its simpler variants, on smaller categories, which are often the most crucial and the hardest to generalize about.

We also introduced the publicly available Drug–Disease Relations dataset, which contains a large training set of crowdsourced labels and a smaller test set of gold labels assigned by experts. This dataset can be used to learn facts about drug–disease relationships that have medical significance, and it shows that combined models like ours can learn effectively in noisy settings.

References

Kevin Atkinson. 2018. SCOWL (Spell Checker Oriented Word Lists). Accessed: September, 2018.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. ArXiv:1312.3005.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, DLRS 2016, pages 7–10, New York, NY, USA. ACM.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18(1):368.

Doug Cutting. 1999. Apache Lucene. Accessed: September, 2018.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, pages 20–28.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784. Association for Computational Linguistics.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895. Association for Computational Linguistics.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. ArXiv:1207.0580.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. ArXiv 1508.01991.

Abhyuday Jagannatha and Hong Yu. 2016. Structured prediction models for RNN based sequence labeling in clinical text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 856–865. Association for Computational Linguistics.

Warren A Kibbe, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, Christopher J Mungall, Janos X Binder, James Malone, Drashtti Vasant, et al. 2014. Disease Ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*, 43(D1):D1071–D1078.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. ArXiv 1412.6980.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Ms, OpenAI.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3567–3575. Curran Associates, Inc.

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.

Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. 2011. Disease Ontology: A backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946.

Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. SNOMED RT: A reference terminology for health care. In *Proceedings of the AMIA Annual Fall Symposium*, page 640. American Medical Informatics Association.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Charles Sutton and Andrew McCallum. 2011. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7(1):92.

Ozlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association : JAMIA*, 14(5):550–563.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Haohan Wang, Xiang Liu, Yifeng Tao, Wenting Ye, Qiao Jin, William W Cohen, and Eric P Xing. 2019. Automatic human-like mining and constructing reliable genetic association database with deep reinforcement learning. *Pacific Symposium on Biocomputing*.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, page bty869.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015. Overview of the BioCreative V chemical disease relation (CDR) task. In *Proceedings of the 5th BioCreative Challenge Evaluation Workshop*, pages 154–166.

Sentence	Hand-built features of word ***bacteria***
antiseptic	**Adjacent words features**:
handwash	word-4:antiseptic, word-3:handwash, word-2:to, word-1:decrease,
to	word:bacteria, word+1:on, word+2:the, word+3:skin, word+4:..
decrease	**Adjacent POS tags features**:
bacteria	tag-4:JJ, tag-3:NN, tag-2:TO, tag-1:VB,
on	tag:NNS, tag+1:IN, tag+2:DT, tag+3:NN, tag+4:..
the	**Semantic environment features**:
skin	bias:1, is_upper:0, is_title:0, is_punctuation:0,
.	in_left_context_of_negative_cues:0, in_right_context_of_negative_cues:0,
	in_left_context_of_prevents_cues:0, in_right_context_of_prevents_cues:0,
	in_left_context_of_treats_cues:0, in_right_context_of_treats_cues:0,
	in_left_context_of_treats_symptoms_cues:0, in_right_context_of_treats_symptoms_cues:0,
	in_left_context_of_contraindicated_cues:0, in_right_context_of_contraindicated_cues:0,
	in_left_context_of_affliction_adj_cues:0, in_right_context_of_affliction_adj_cues:0,
	in_left_context_of_indication_cues:0, in_right_context_of_indication_cues:0,
	in_left_context_of_details_cues:0, in_right_context_of_details_cues:0.

Table A1: Hand-built features of the word ***bacteria*** in a Drug–Disease Relations dataset example. These features describe the word's adjacent words, adjacent POS tags, and semantic environment (section 2). The detailed meanings of hand-built features in the table are described as below: **Adjacent words features**: "word($\pm$1/2/3/4)" feature the word and adjacent words within a window size of 9. **Adjacent POS tags features**: "tag($\pm$1/2/3/4)" feature the tags of word and its adjacent words within a window size of 9. **Semantic environment features**: "bias" is always 1 for all words; "is_upper" specifies whether the word is upper case or lower case; "is_title" features whether the word is in the title or not; "is_punctuation" specifies whether the token is actually a word or a punctuation. "in_left/right_context_of_negative/prevents/treats(_symptoms)/contraindicted/afflicition_adj/indication/details_cues" feature whether the word is in the left or right context (of specific window size like 4) of cue-words from specific lexicons. Features related to 8 lexicons are shown in this example. Concrete examples: *not*, *none* and *no* are three cue-words of lexicon "negative_cues", *prevent* and *avoid* are two cue-words of lexicon "prevents_cues", *treat*, *solve* and *alleviate* are three cue-words of lexicon "treats_cues" etc. Different semantic environments are defined in the five datasets by carefully defining the lexicons/cue-words from various sources which possibly contain corresponding domain knowledge, as discussed in section 2 and section 3.

Statistics	Diagnosis Detection	Prescription Reasons	Penn Adverse Drug Reactions (ADR)	Chemical–Disease Relations (CDR)	Drug–Disease Relations
# texts	6042	5179	–	–	–
# training texts	–	–	749	1000	9494
# test texts	–	–	272	500	500
mean text length	17	19	19	227	30
max text length	374	258	40	623	542
# labels	4	4	5	3	5

Table A2: Statistics for our five datasets. The sample size varies from around 1,000 to 10,000. The mean text length (measured as the number of words) varies from 17 (short sentences) to 227 (full paragraphs). The number of labels varies from 3 to 5. ADR, CDR, and Drug–Disease Relations are already partitioned into training and test sets, while Diagnosis Detection and Prescription Reasons do not have predefined splits.

Models	Hyperparams	Diagnosis Detection	Prescription Reasons	Penn Adverse Drug Reactions (ADR)	Chemical–Disease Relations (CDR)	Drug–Disease Relations
rand-LSTM-CRF	η	1e-4	1e-4	1e-4	1e-4	1e-4
	$\text{epoch}_{\text{tune}}$	3	3	513	10	13
	$\text{epoch}_{\text{train}}$	34	40	3076	164	130
	$\mathcal{R}_{c1}$			{ 0, 3e-5, 1e-4, 3e-4, 1e-3 }		
	$\mathcal{R}_{c2}$			{ 0, 3e-4, 1e-3, 3e-3, 1e-2 }		
HB-CRF	η	1e-2	1e-2	3e-2	1e-2	1e-4
	$\text{epoch}_{\text{tune}}$	1	1	10	2	3
	$\text{epoch}_{\text{train}}$	3	4	82	10	35
	$\mathcal{R}_{c1}$			{ 0, 3e-6, 1e-5, 3e-5, 1e-4 }		
	$\mathcal{R}_{c2}$			{ 0, 3e-5, 1e-4, 3e-4, 1e-3 }		
ELMo-LSTM-CRF	η	1e-3	1e-3	1e-4	1e-3	5e-6
	$\text{epoch}_{\text{tune}}$	1	1	10	2	3
	$\text{epoch}_{\text{train}}$	3	4	82	10	35
	$\mathcal{R}_{c1}$			{ 0, 3e-5, 1e-4, 3e-4, 1e-3 }		
	$\mathcal{R}_{c2}$			{ 0, 3e-4, 1e-3, 3e-3, 1e-2 }		
ELMo-LSTM-CRF-HB	η	1e-3	1e-3	1e-4	1e-3	1e-5
	$\text{epoch}_{\text{tune}}$	1	1	10	2	3
	$\text{epoch}_{\text{train}}$	3	4	82	5	35
	$\mathcal{R}_{c1}$			{ 0, 3e-7, 1e-6, 3e-6, 1e-5 }		
	$\mathcal{R}_{c2}$			{ 0, 3e-6, 1e-5, 3e-5, 1e-4 }		

Table A3: Hyperparameters for our experiments. The step size η is first manually tuned within the training set when the ℓ_1 and ℓ_2 regularizers are set to be zeros. The coefficients c_1 and c_2 of the ℓ_1 and ℓ_2 regularizers are determined via random search (for 10 settings) from ranges $\mathcal{R}_{c1}$ and $\mathcal{R}_{c2}$ during tuning (Bergstra and Bengio, 2012). Epochs of tuning $\text{epoch}_{\text{tune}}$ are set to 1$\sim$3 to reduce tuning time for most datasets (which consumes most of the time for the experiments). It is set to 10 for ADR since that dataset is so small that it is hard to see clear trends after just one epoch. Epochs of training $\text{epoch}_{\text{train}}$ are set to be large enough until the training converges. The 'rand-LSTM-CRF' model requires many more epochs for tuning and training because of the updates to the randomly initialized embeddings.

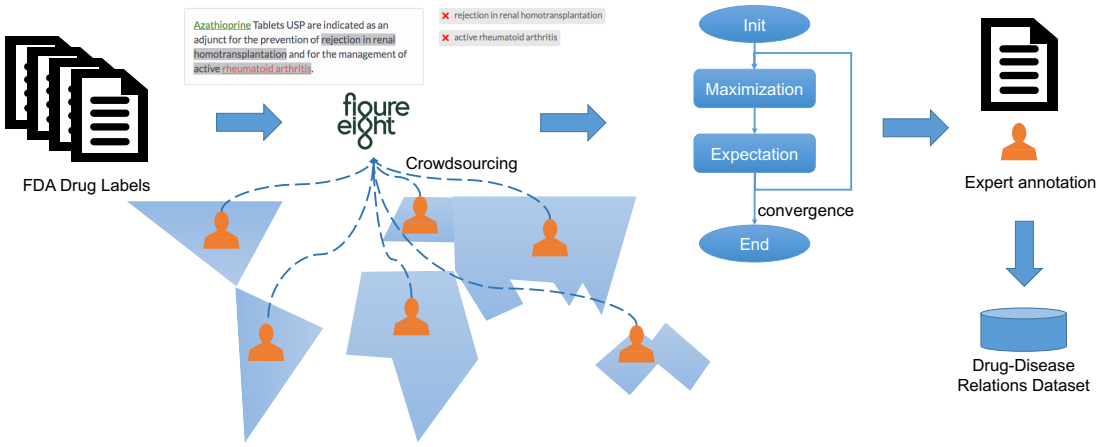

Figure A1: Procedure for building the Drug–Disease Relations dataset. 10,000 raw sentences from the FDA Drug Labels corpus were annotated by participants from 72 countries on the Figure Eight platform (crowdsourcing). Expectation Maximization was used to infer labels for all the annotated sentences used for training. A team of experts independently labeled different examples for testing. The resulting dataset consists of 9,500 crowdsourced examples and 500 expert-annotated examples.

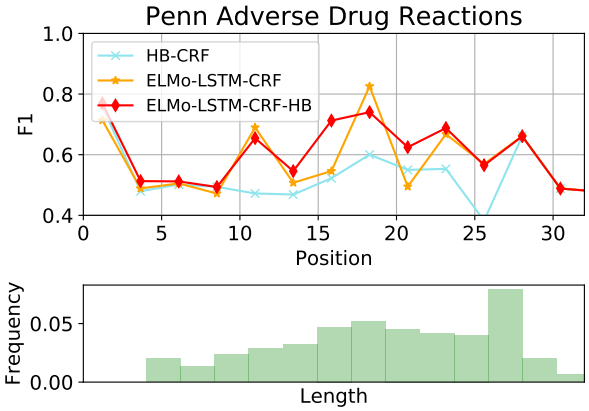

Figure A2: Text-length experiment for the Penn Adverse Drug Reactions (ADR) dataset. Since ADR uses the IOB tag format, in order to calculate per-token F1 scores, we collapse test-set labels starting with 'B-' and 'I-' into the same labels. The ELMo-LSTM-CRF always performs better than the HB-CRF, while the combined model takes advantage of both models and always outperforms both base models. Figure 2 provides comparable plots for the other four datasets.

Penn Adverse Drug Reactions (ADR)			
Label	Support	F1 score	Improvement
OTHER	5023	98.0	0.3%
ADR	283	57.1	17.7%
INDICATION	29	35.9	178.3%

Table A4: Relative F1 score improvements of different labels in the Penn Adverse Drug Reactions (ADR) dataset. To bring the IOB tag format of this dataset in line with our others, ADR merges B-ADR and I-ADR, and INDICATION merges B-INDICATION and I-INDICATION. Consistent with table 3, the combined model gains most in the smallest categories.