Effective Feature Representation for Clinical Text Concept Extraction

Yifeng Tao[1,2], Bruno Godefroy[1], Guillaume Genthial[1], Christopher Potts[1,3,*]

[1]Roam Analytics

[2]Carnegie Mellon University

[3]Stanford University

Background: Healthcare Text Datasets

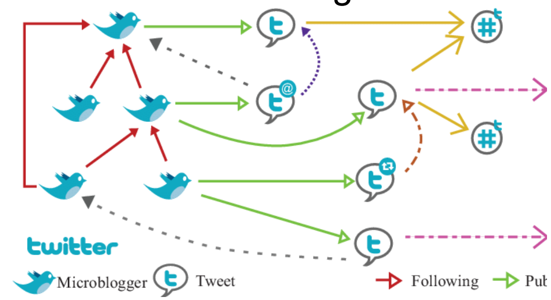
- Crucial information of healthcare recorded only in free-form text

Scientific
Chemical-Disease Relations

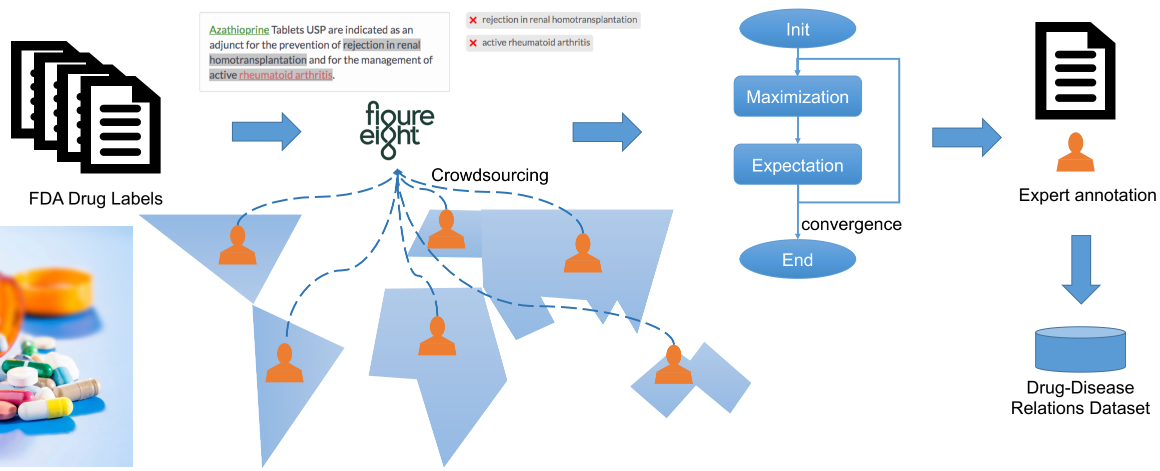
PubMed
NCBI

Clinical
Diagnosis Detection
Prescription Reasons

Social
Penn Adverse Drug Reactions



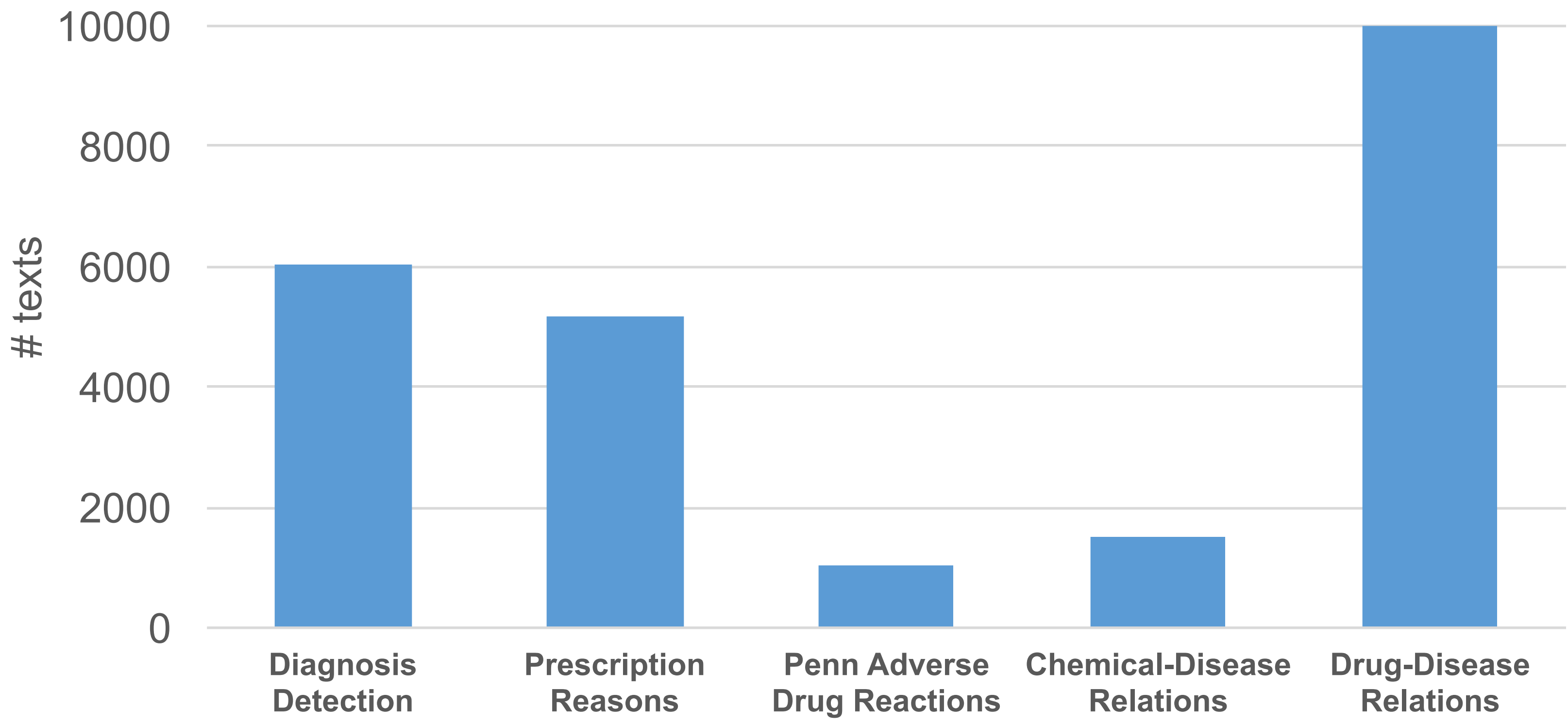
Commercial
Drug-Disease Relations



[Figures from: 1. Lamjed Ben Jabeur et al. Uprising microblogs: A Bayesian network retrieval model for tweet search. 2012, 2. https://www.sjm.com.br/utilidades/pubmed-busca, 3. http://anakin.uta.cloud/uncategorized/the-need-for-drug-donations, 4. https://www.autismawareness.com.au/news-events/aupdate/is-there-an-over-diagnosis-of-autism]

Background: Healthcare Text Datasets

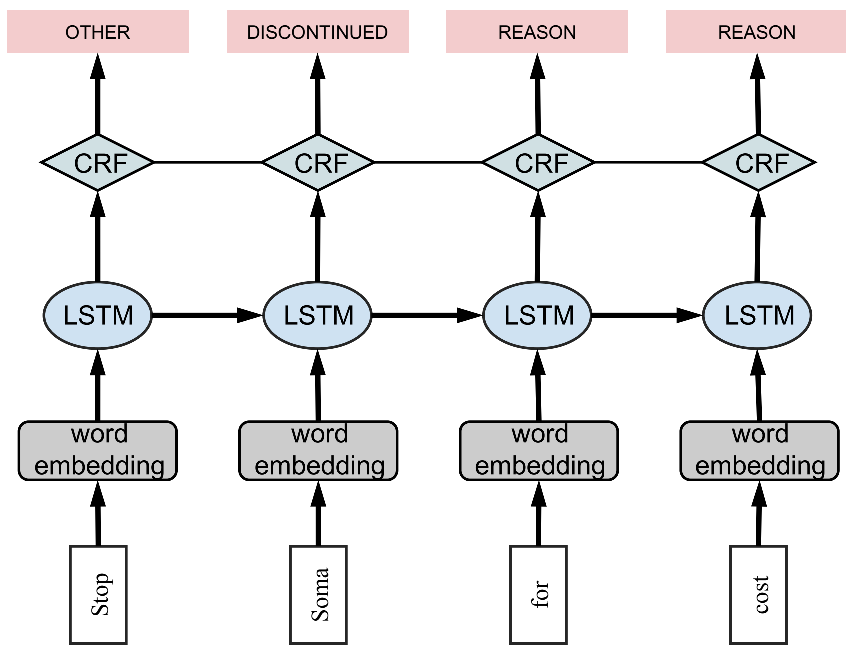
- Clinical text datasets are scarce and expensive
  - Privacy considerations
  - Domain specialists

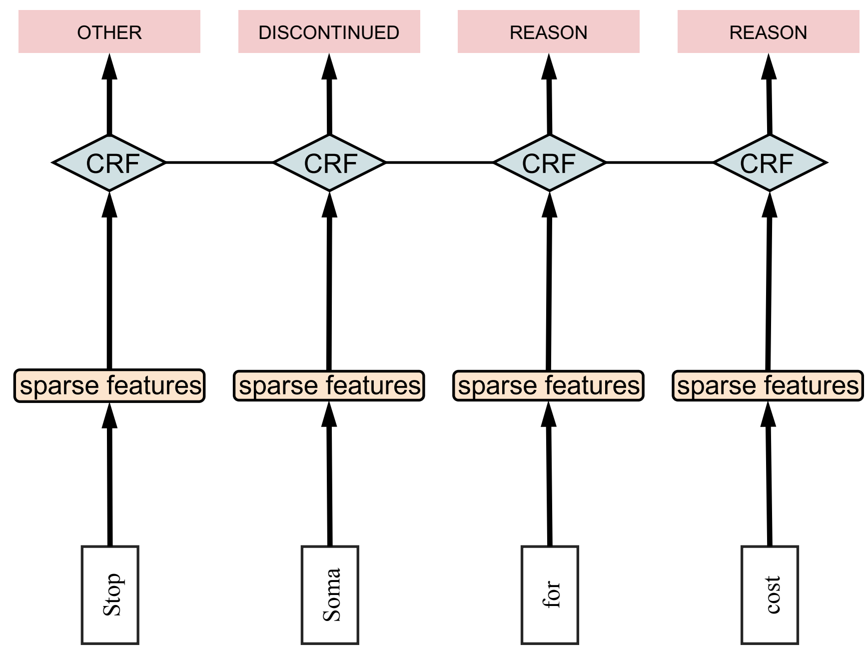


Task: Clinical Text Annotation

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

Previous Models

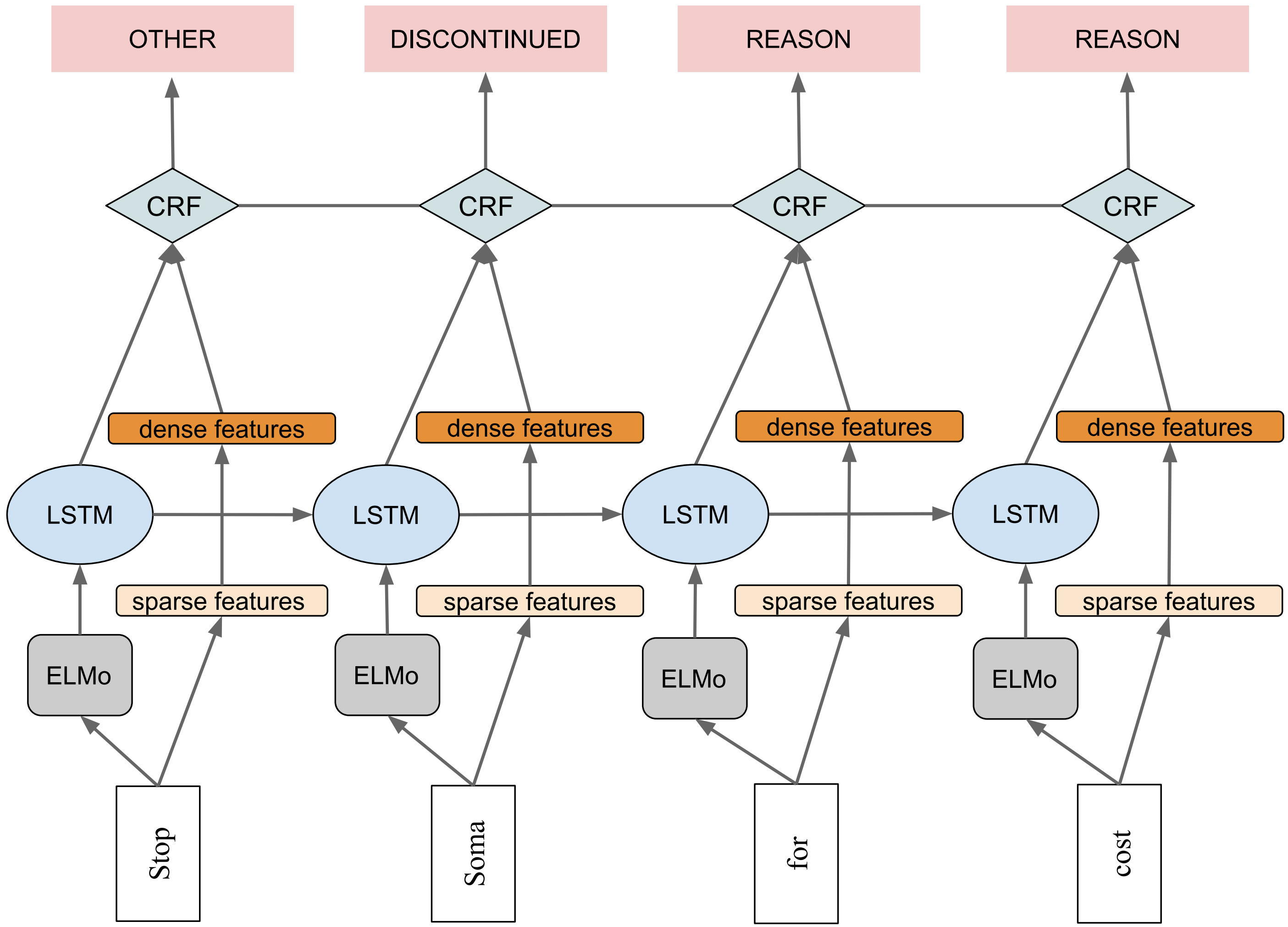




- LSTM-CRF
  - General text
  - Distributed word embeddings
- HB-CRF
  - Clinical text
  - Sparse hand-built features

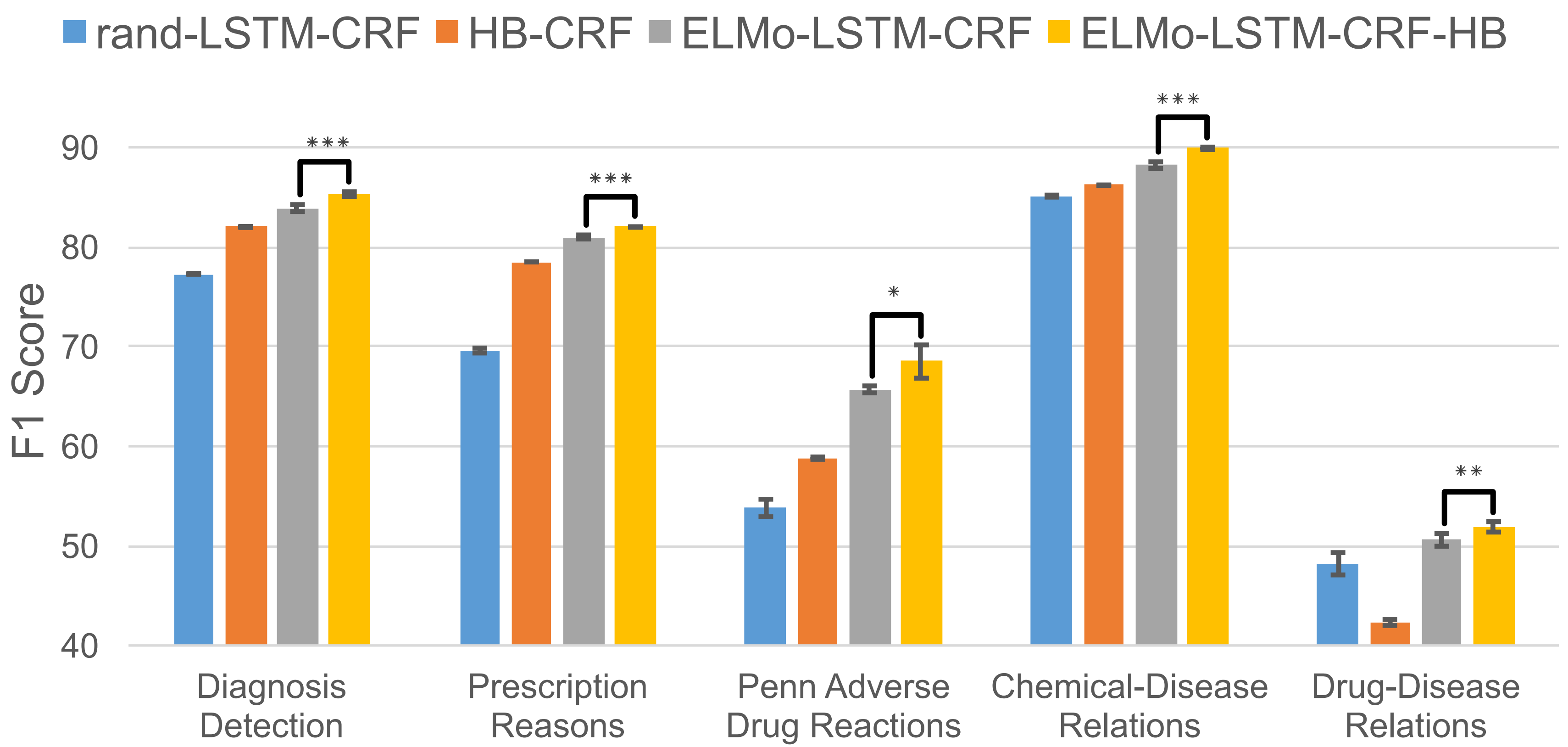
Model: ELMo-LSTM-CRF-HB

- Dense **ELMo** word embeddings + Sparse **hand-built** features



Performance: Per-token Macro-F1 Scores

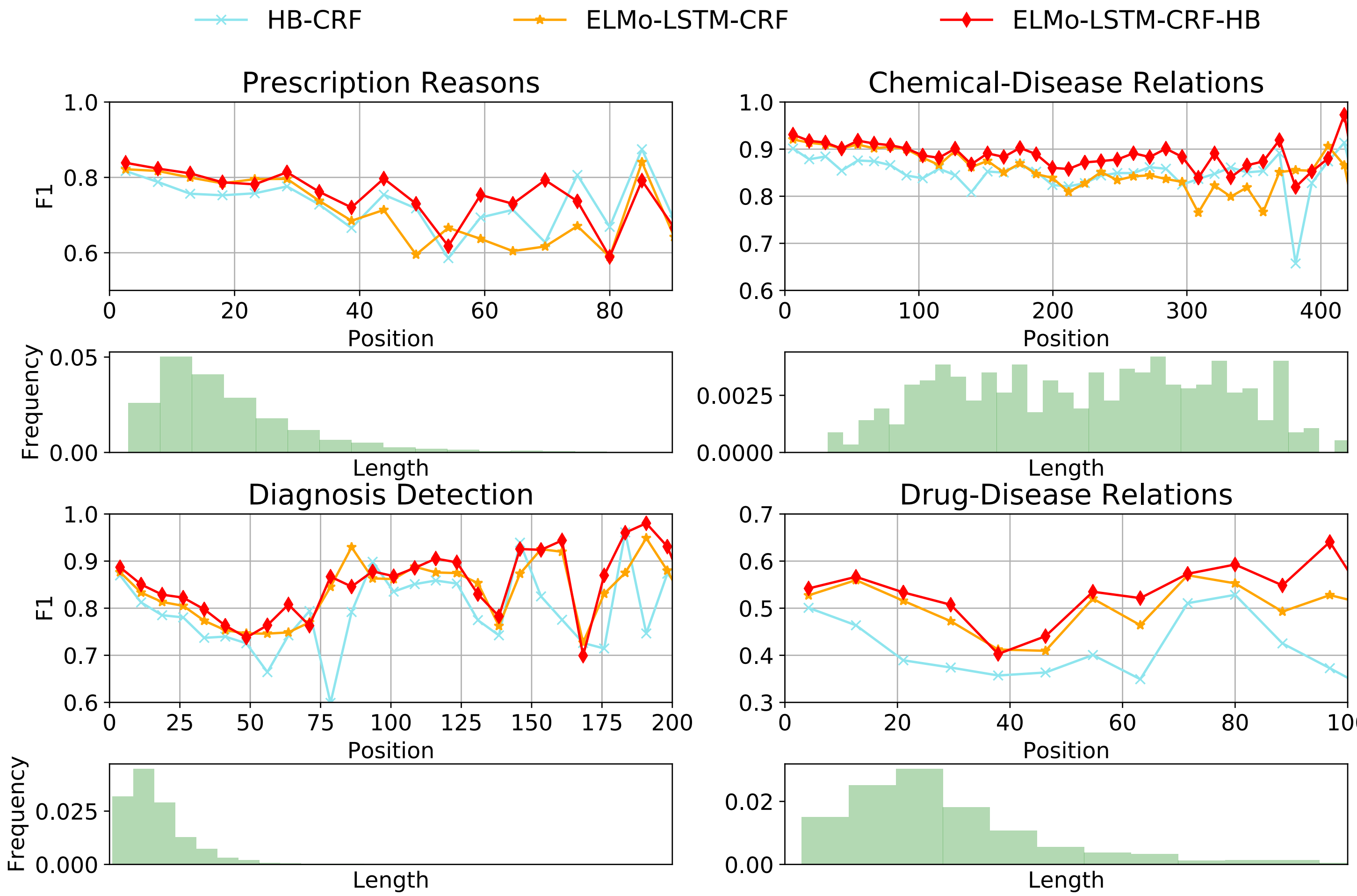
- Hyperparameters tuned through cross-validation
- Each experiment repeated for five times



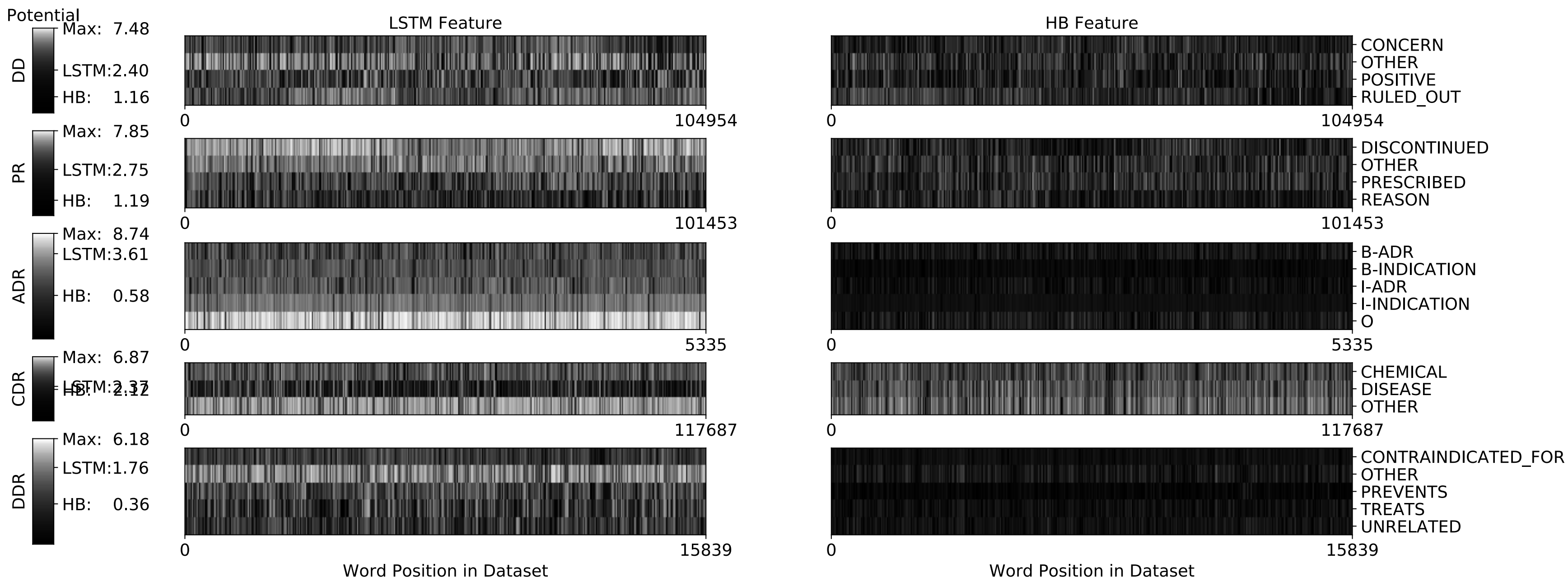
*: p<0.05, **: p<0.01, ***: p<0.001

The Role of Text Length

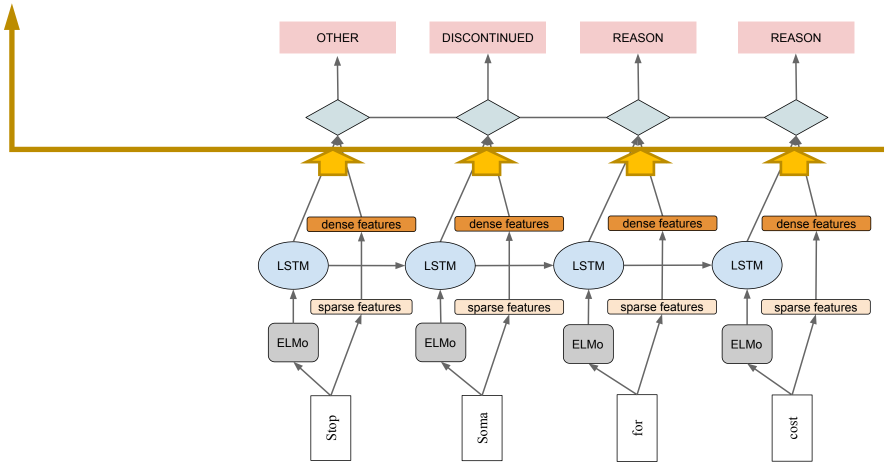
- LSTM: handles short texts well
- HB-CRF: robust on long texts



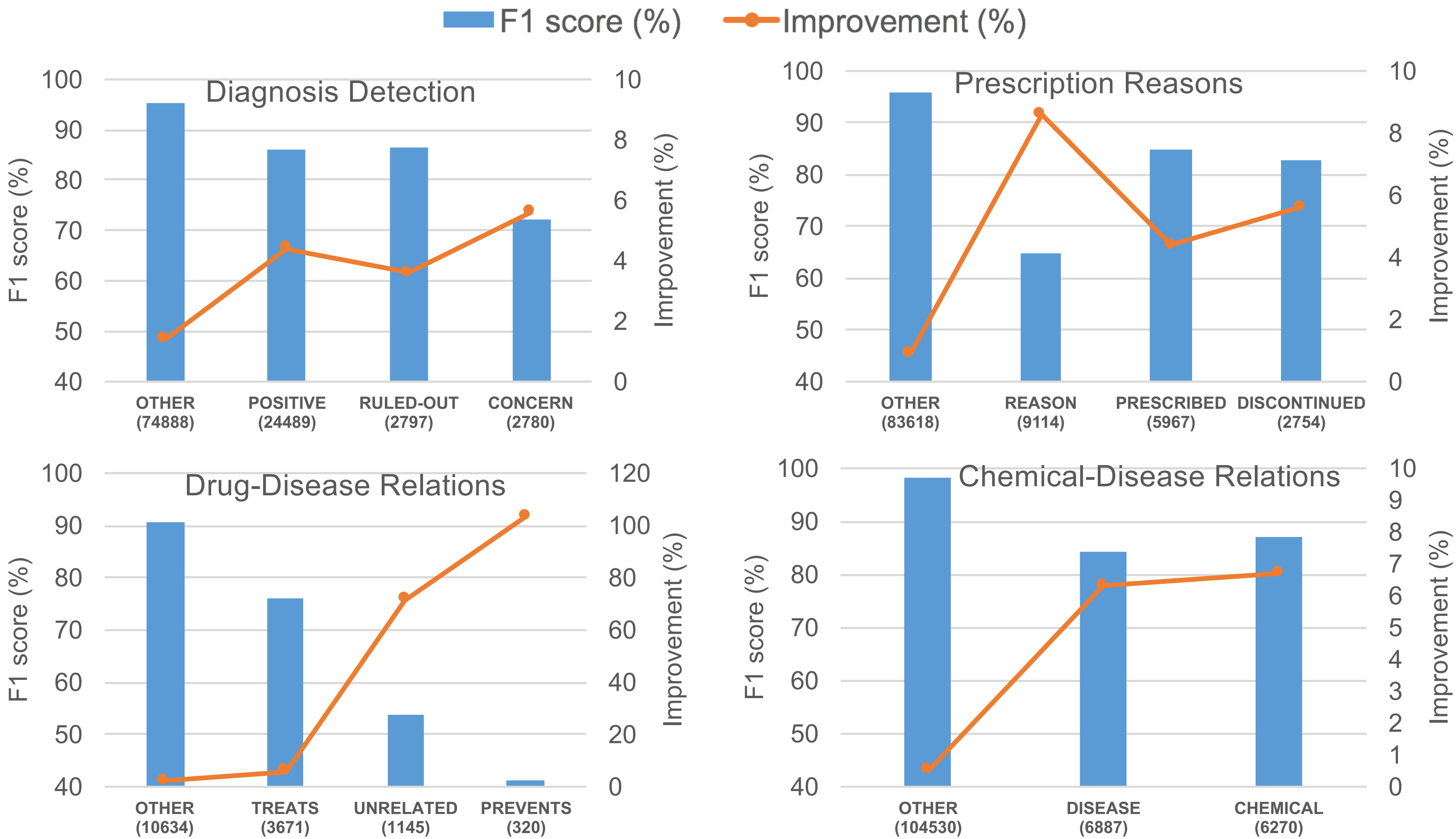
CRF Potential Scores



- LSTM features always more important
- HB features make substantial contribution



Major Improvements in Minor Categories



Label/Category (Support)

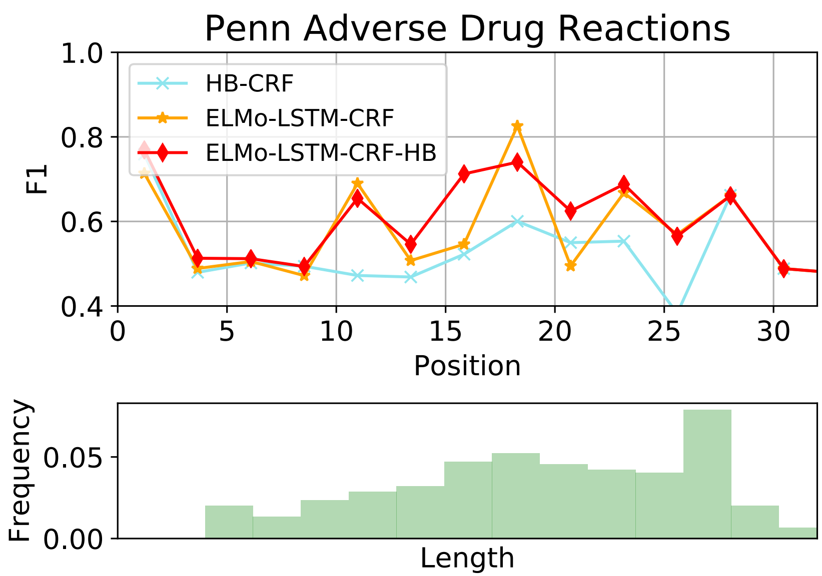
Conclusion

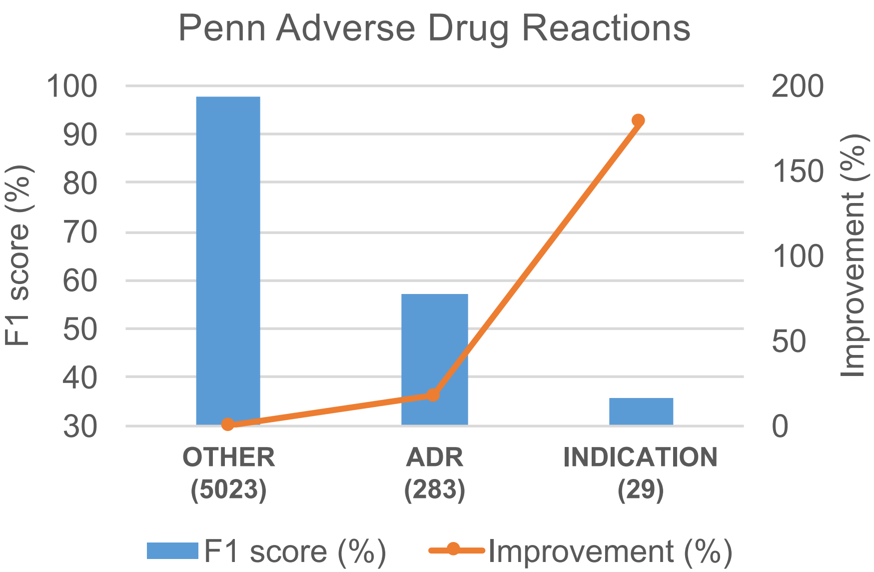
- A unified feature representation for clinical text sequence labeling
  - Sparse, ontology-driven features
  - Dense LSTM features
- Best performance on five distinct healthcare datasets
  - Takes advantages of both feature types
  - Makes maximal use of small, expensive, domain-specific healthcare texts
- A new labeled clinical dataset
  - Identifies the treatment relations between drugs and diseases
- Extensive analysis to identify what information our model makes use of, and why its performance is consistently improved

Acknowledgement

- Roam Analytics
  - Christopher Potts
  - Bruno Godefroy
  - Guillaume Genthial
  - Kevin Reschke
  - NLP Group

Penn Adverse Drug Reactions (ADR) Results



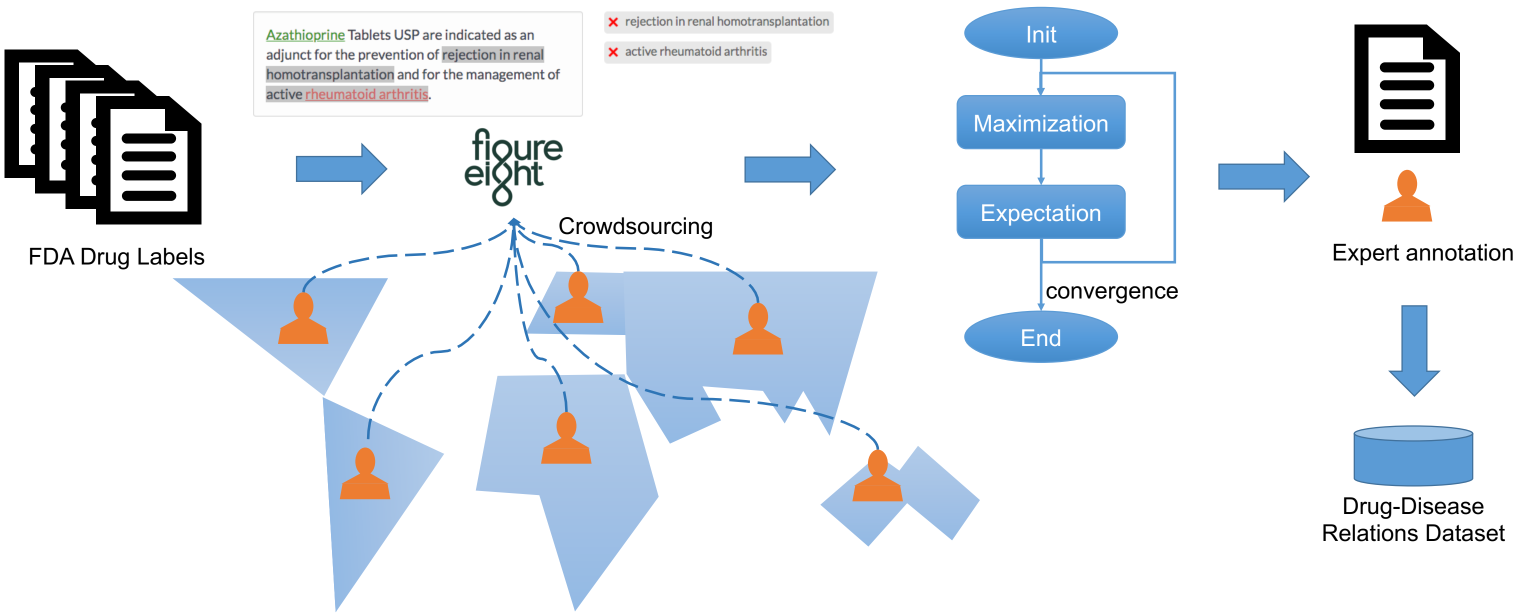


- The Role of Text Length
- Major Improvements in Minor Categories

Example of Hand-built Features

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.

Procedure for Building Drug-Disease Relations Dataset



Statistics of Datasets

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

Hyperparameters of Experiments

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 }				