MeasEval: Measurement Extraction from Scientific Texts

Literature Review

A. Furkan Okuyucu

Overview

- Brief Task Description
- Sequence Labeling
- Taxonomy of Approaches
- Rule-based models(2)
- biLSTM Models(2)
- Transformer applications(4)

Task Description

MeasEval

- **Input:** The *soda can*'s volume was 355 ml after I drank half the can.
- **Output:**
  - Quantity = 335 ml
  - Measured Entity = soda can
  - Measured Property = volume
  - Qualifier = after I drank half the can

More examples for quantity spans

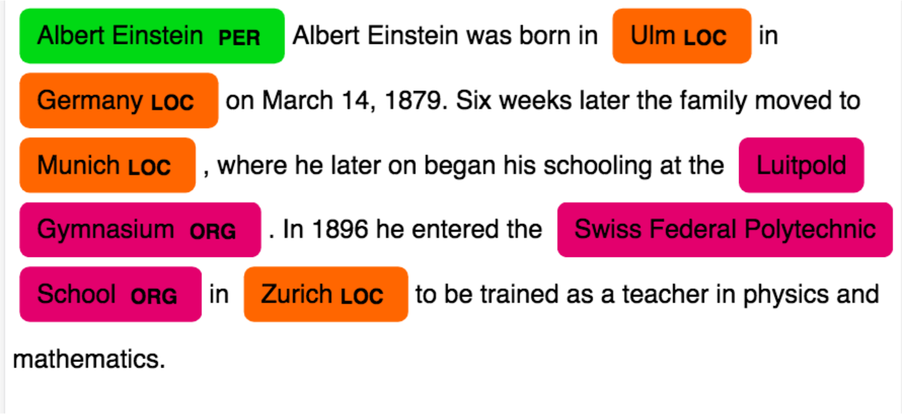
- **around 1300 m s−1**
- **four** transits
- **range of 1.5–2.6 m**
- **4.5 kg, 6 kg and 13 kg**
- Standard Deviation
    - **2SD of 1.23±0.25‰**
- Tolerance
    - **9 ± 6%.**

Sequence Labeling

- Task of pattern recognition
- Labeling group of morphemes according to task
- Some subparts

1. Part of Speech Tag(POS)

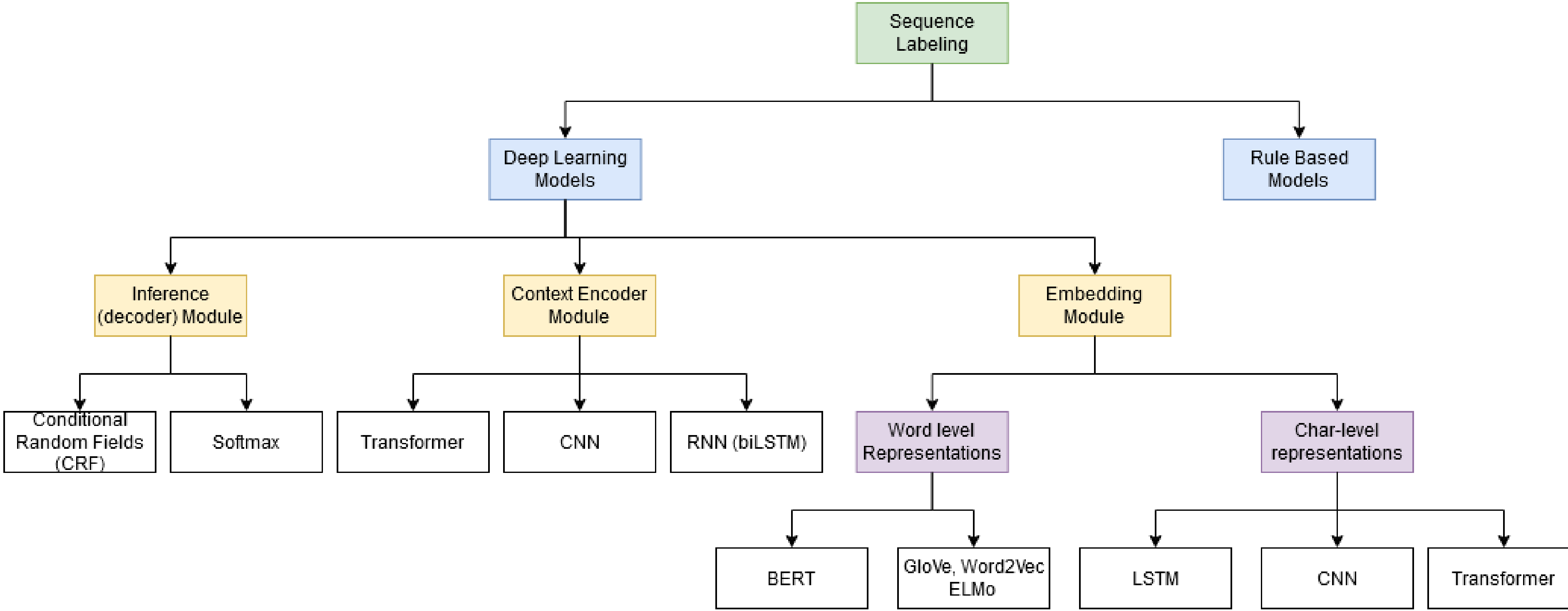
2. Named Entity Recognition(NER)



Evaluation Metrics and Datasets

- Most used datasets
  - POS
    - Wall Street Journal (WSJ)
  - NER
    - CoNLL 2003
- Metrics
  - POS
    - Accuracy
  - NER
    - F1 score

Sequence Labeling Models



How to extract unit of measures in scientific texts?(2013) [1]

Approach
- Locate units with supervised Learning
- Use string distance to extract units

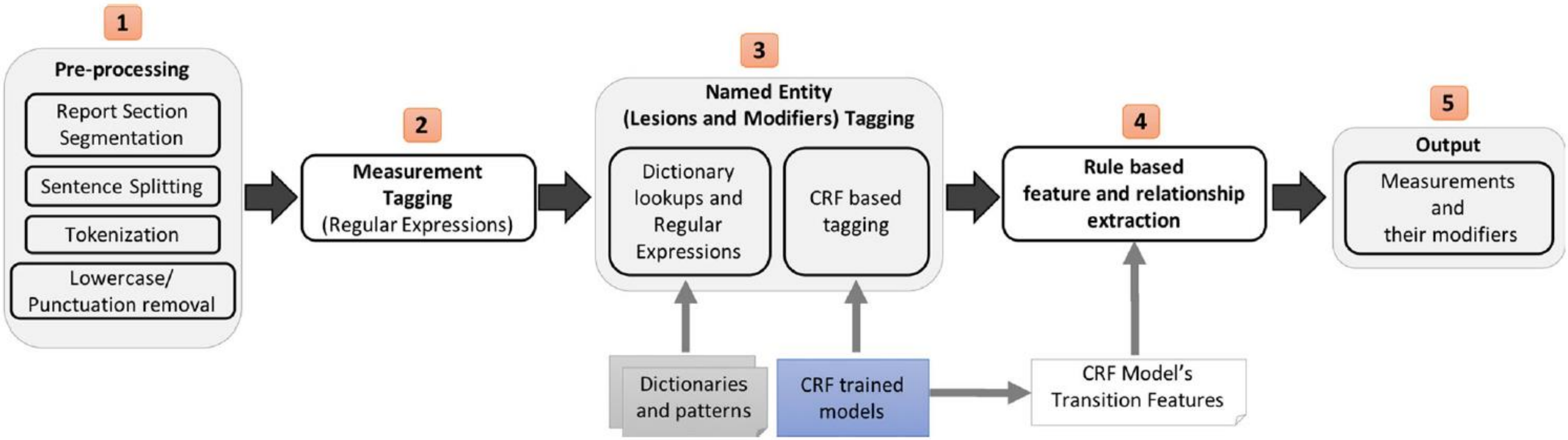
Data
- 35000 sentences from food science domain

Results
- Recall =0.95

Problems
- Limited to one area
- Requires handcrafted features, gazettes

Automated Detection of Measurements and Their Descriptors in Radiology Reports [2]

Motivation	Radiological reports are in free text format Hard to extract measurements for treatment planning
Approach	CRF rule based feature extraction
Data	1117 CT/MR training 100 CT/MR test
Results	F1 = 98.18
Problems	Domain Specific Handcrafted Features Not Generalizable



Proposed Pipeline

Neural Architectures for Named Entity Recognition[3]

Motivation

Neural architecture

no language specific resource and features

Approach

Char Embedding (biLSTM)

Word Embeddings (Word2Vec)

BiLSTM

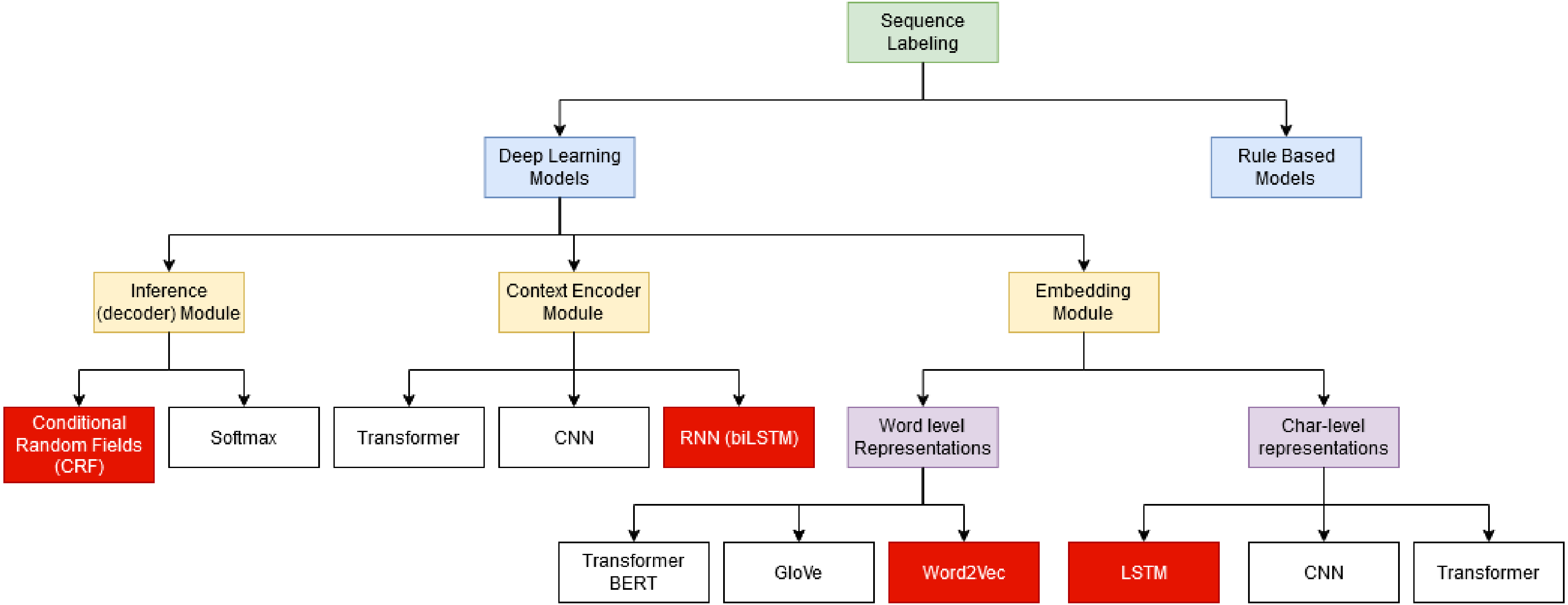
CRF

Data

CoNLL-2003

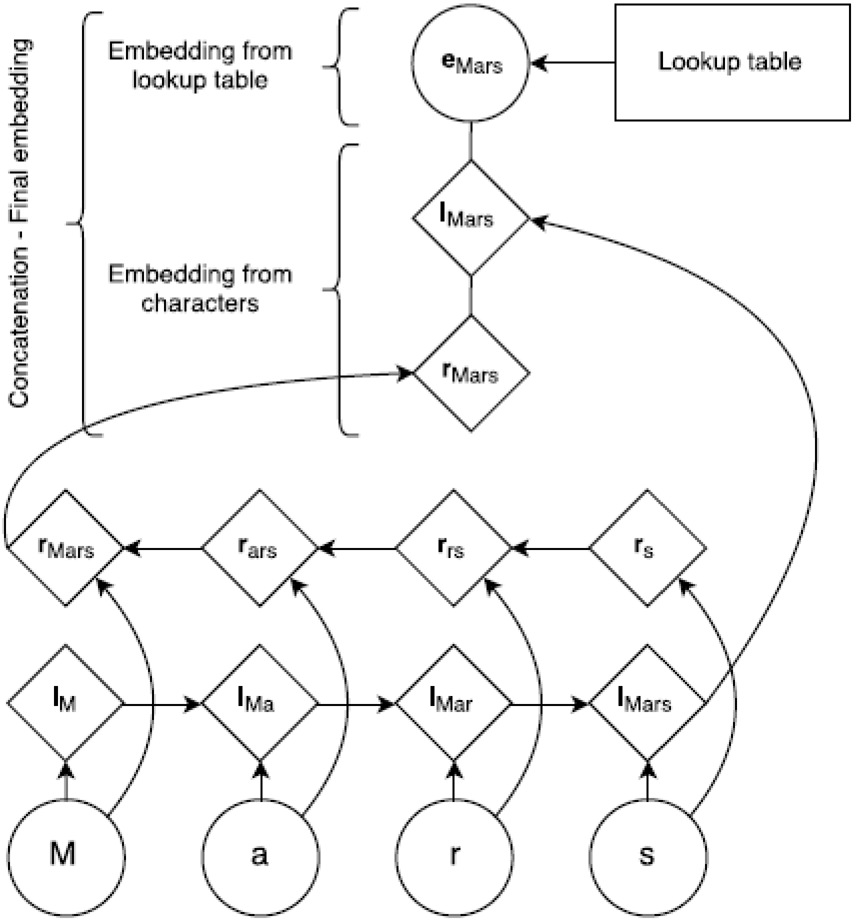
Results

F1 =90.94

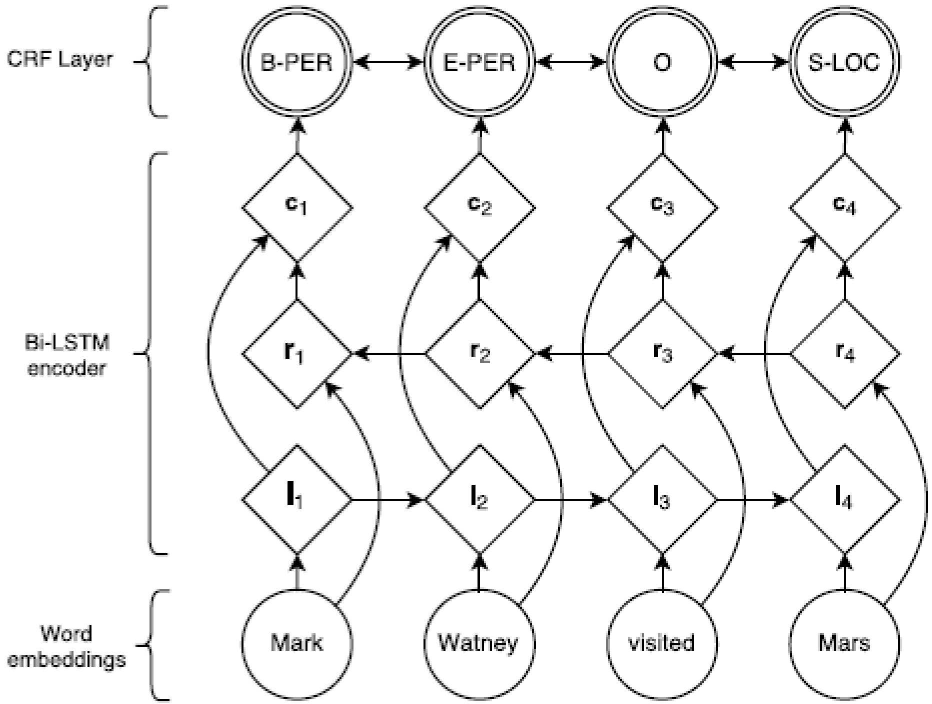


Sequence Labeling
Deep Learning Models
Rule Based Models
Inference (decoder) Module
Context Encoder Module
Embedding Module
Conditional Random Fields (CRF)
Softmax
Transformer
CNN
RNN (biLSTM)
Word level Representations
Char-level representations
Transformer BERT
GloVe
Word2Vec
LSTM
CNN
Transformer

Character Embeddings (biLSTM)



Main Architecture



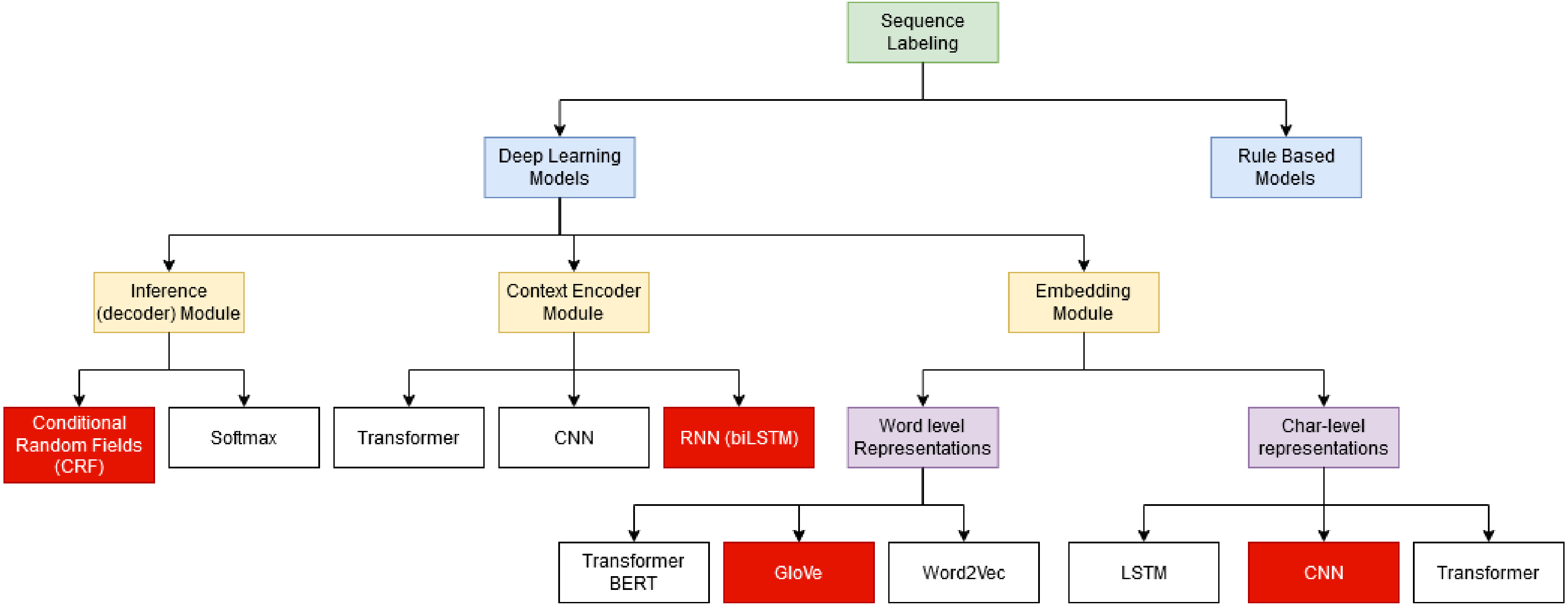
Results

Model	F_1
Collobert et al. (2011)*	89.59
Lin and Wu (2009)	83.78
Lin and Wu (2009)*	90.90
Huang et al. (2015)*	90.10
Passos et al. (2014)	90.05
Passos et al. (2014)*	90.90
Luo et al. (2015)* + gaz	89.9
Luo et al. (2015)* + gaz + linking	**91.2**
Chiu and Nichols (2015)	90.69
Chiu and Nichols (2015)*	90.77
LSTM-CRF (no char)	90.20
LSTM-CRF	**90.94**
S-LSTM (no char)	87.96
S-LSTM	90.33

Table 1: English NER results (CoNLL-2003 test set). * indicates models trained with the use of external labeled data

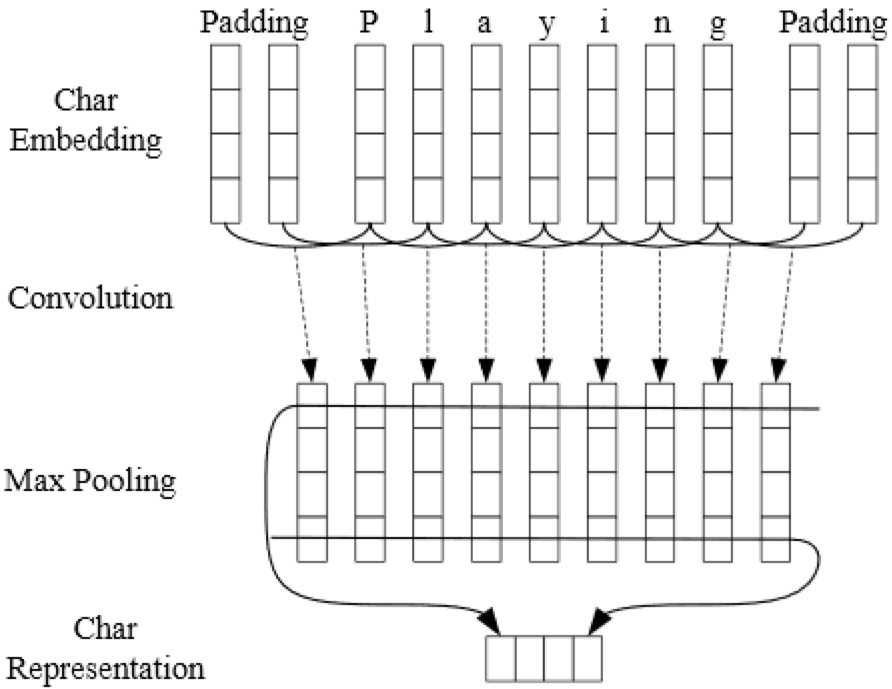
End to end Sequence Labeling Via Bidirectional LSTM-CNNs-CRF [4]

Motivation	no task specific resource no data processing no feature engineering
Approach	Char Embedding (CNN) Word Embeddings (GloVe) BiLSTM CRF
Data	CoNLL-2003
Results	NER F1 =91.21 POS Accuracy = 97.55

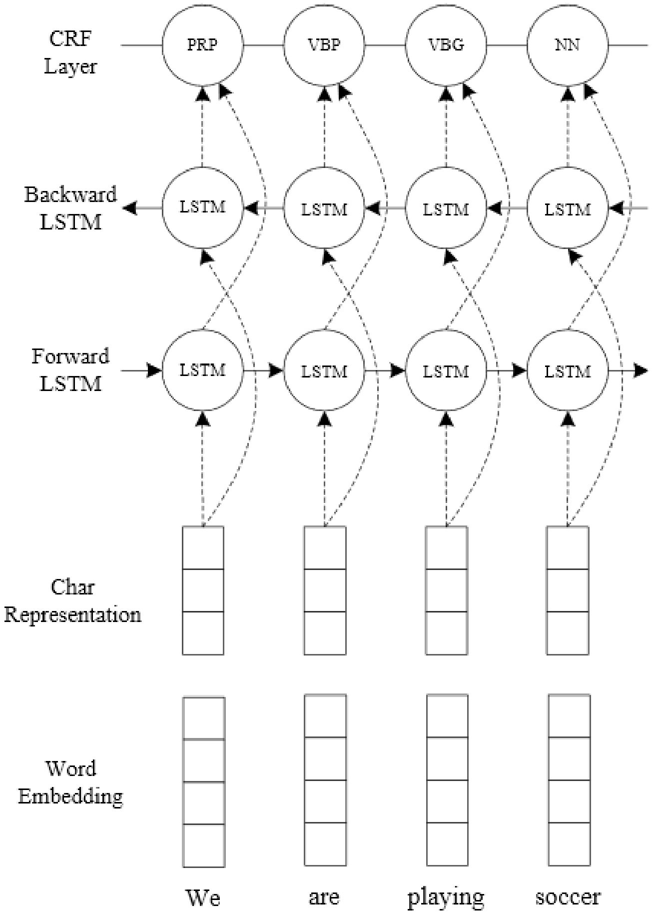


Sequence Labeling
Deep Learning Models
Rule Based Models
Inference (decoder) Module
Context Encoder Module
Embedding Module
Conditional Random Fields (CRF)
Softmax
Transformer
CNN
RNN (biLSTM)
Word level Representations
Char-level representations
Transformer BERT
GloVe
Word2Vec
LSTM
CNN
Transformer

Character Embedding (CNN)



Main Architecture



Hyperparameters

Layer	Hyper-parameter	POS	NER
CNN	window size	3	3
	number of filters	30	30
LSTM	state size	200	200
	initial state	0.0	0.0
	peepholes	no	no
Dropout	dropout rate	0.5	0.5
	batch size	10	10
	initial learning rate	0.01	0.015
	decay rate	0.05	0.05
	gradient clipping	5.0	5.0

Table 1: Hyper-parameters for all experiments.

Dataset

Dataset		WSJ	CoNLL2003
Train	SENT	38,219	14,987
	TOKEN	912,344	204,567
Dev	SENT	5,527	3,466
	TOKEN	131,768	51,578
Test	SENT	5,462	3,684
	TOKEN	129,654	46,666

Table 2: Corpora statistics. SENT and TOKEN refer to the number of sentences and tokens in each data set.

Results for NER

Model	F1
Chieu and Ng (2002)	88.31
Florian et al. (2003)	88.76
Ando and Zhang (2005)	89.31
Collobert et al. (2011)‡	89.59
Huang et al. (2015)‡	90.10
Chiu and Nichols (2015)‡	90.77
Ratinov and Roth (2009)	90.80
Lin and Wu (2009)	90.90
Passos et al. (2014)	90.90
Lample et al. (2016)‡	90.94
Luo et al. (2015)	91.20
This paper	**91.21**

- Reference to 3

Results for POS

Model	**Acc.**
Giménez and Màrquez (2004)	97.16
Toutanova et al. (2003)	97.27
Manning (2011)	97.28
Collobert et al. (2011)‡	97.29
Santos and Zadrozny (2014)‡	97.32
Shen et al. (2007)	97.33
Sun (2014)	97.36
Søgaard (2011)	97.50
This paper	**97.55**

Table 4: POS tagging accuracy of our model on test data from WSJ proportion of PTB, together with top-performance systems. The neural network based models are marked with ‡.

Star-Transformer

Motivation

Decreasing the complexity of Transformer O(n^2)

capturing local composition and long dependencies using attention mechanism

Approach

Char Embedding

Star Transformer

CRF

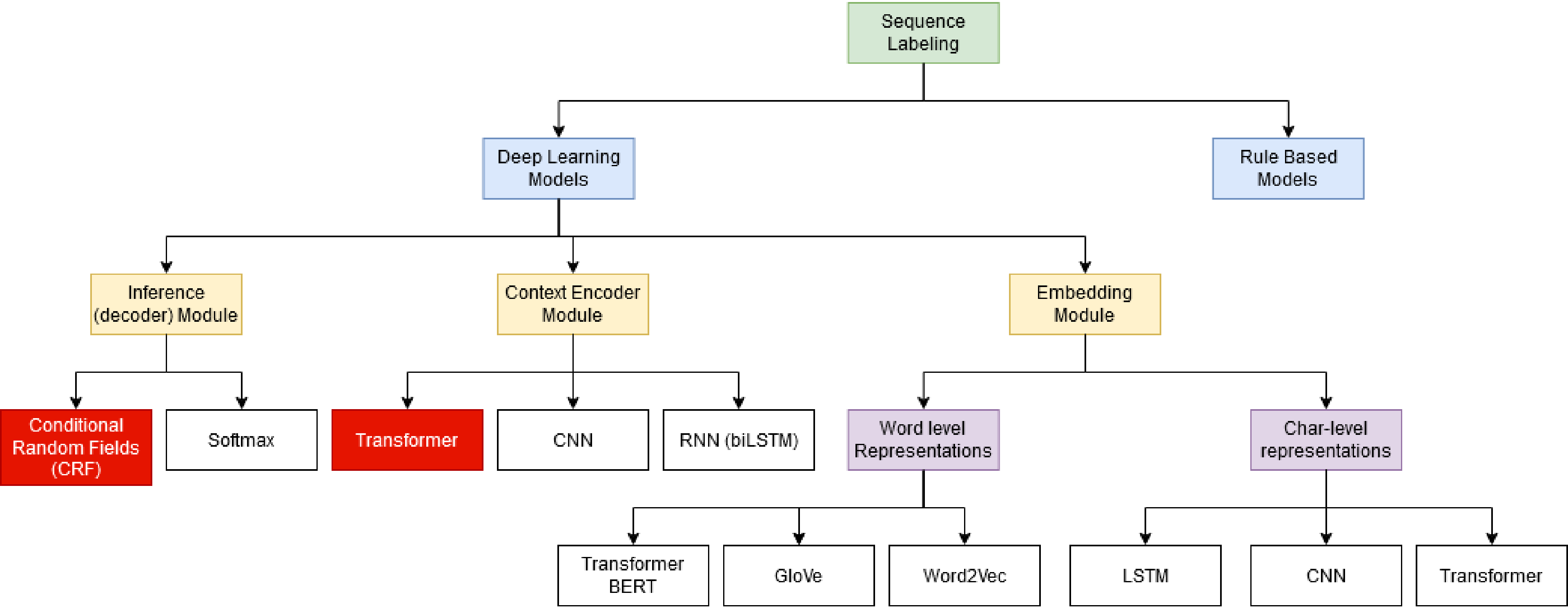
Data

CoNLL-2003

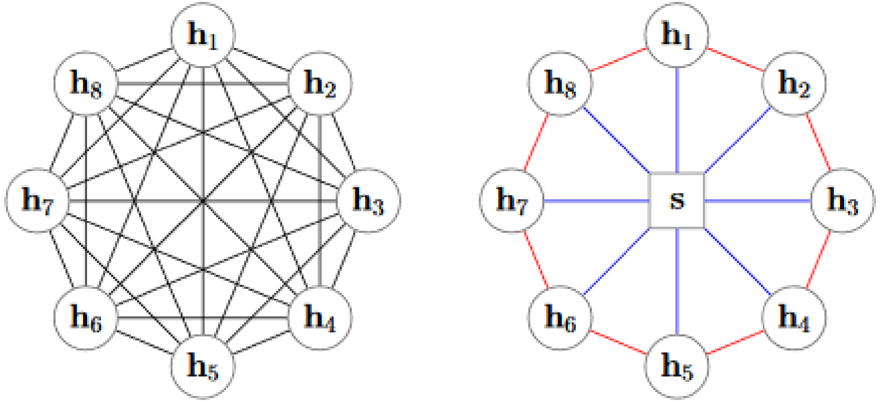
Results

NER F1 =91.98

POS Accuracy = 97.68


Sequence Labeling
Deep Learning Models
Rule Based Models
Inference (decoder) Module
Context Encoder Module
Embedding Module
Conditional Random Fields (CRF)
Softmax
Transformer
CNN
RNN (biLSTM)
Word level Representations
Char-level representations
Transformer BERT
GloVe
Word2Vec
LSTM
CNN
Transformer

Architecture



Virtual relay node
Satellite nodes

Final state of satellite nodes H_T are given to CRF Layer to label the words

Results Comparison

Model	Adv Tech		POS	NER	
			PTB	CoNLL2003	CoNLL2012
	char	CRF	Acc	F1	F1
(Ling et al., 2015)	✓	✓	97.78	-	-
(Collobert et al., 2011)	✓	✓	97.29	89.59	-
(Huang et al., 2015)	✓	✓	97.55	90.10	-
(Chiu and Nichols, 2016a)	✓	✓	-	90.69	86.35
(Ma and Hovy, 2016)	✓	✓	97.55	91.06	-
(Nguyen et al., 2016)	✓	✓	-	91.2	-
(Chiu and Nichols, 2016b)	✓	✓	-	91.62	86.28
(Zhang et al., 2018)	✓	✓	97.55	91.57	-
(Akhundov et al., 2018)	✓	✓	97.43	91.11	87.84
Transformer			96.31	86.48	83.57
Transformer + Char	✓		97.04	88.26	85.14
Star-Transformer			97.14	90.93	86.30
Star-Transformer + Char	✓		97.64	91.89	87.64
Star-Transformer + Char + CRF	✓	✓	**97.68**	**91.98**	**87.88**

- Reference to 4

A Survey on Recent Advances in Sequence Labeling from Deep Learning Models [6]

Method	F1-score
Collobert et al. 2011 [17]	88.67%
Kuru et al. 2016 [50]	84.52%
Chiu and Nichols 2016 [13]	90.91%
Lample et al. 2016 [52]	90.94%
Ma and Hovy 2016 [71]	91.21%
Rei 2017 [91]	86.26%
Strubell et al. 2017 [104]	90.54%
Zhang et al. 2017 [126]	90.70%
Tran et al. 2017 [109]	91.23%
Wang et al. 2017 [113]	91.24%
Sato et al. 2017 [101]	91.28%
Shen et al. 2018 [103]	90.69%
Zhang et al. 2018 [127]	91.22%
Liu et al. 2018 [65]	91.24%
Ye and Ling 2018 [122]	91.38%
Gregoric et al. 2018 [26]	91.48%
Zhang et al. 2018 [128]	91.57%
Xin et al. 2018 [116]	91.64%
Hu et al. 2019 [34]	91.40%
Chen et al. 2019 [10]	91.44%
Yan et al. 2019 [118]	91.45%
Liu et al. 2019 [67]	91.96%
Luo et al. 2020 [68]	91.96%
Jiang et al. 2020 [39]	92.2%
Li et al. 2020 [58]	92.67%

NER task on CoNLL 2003

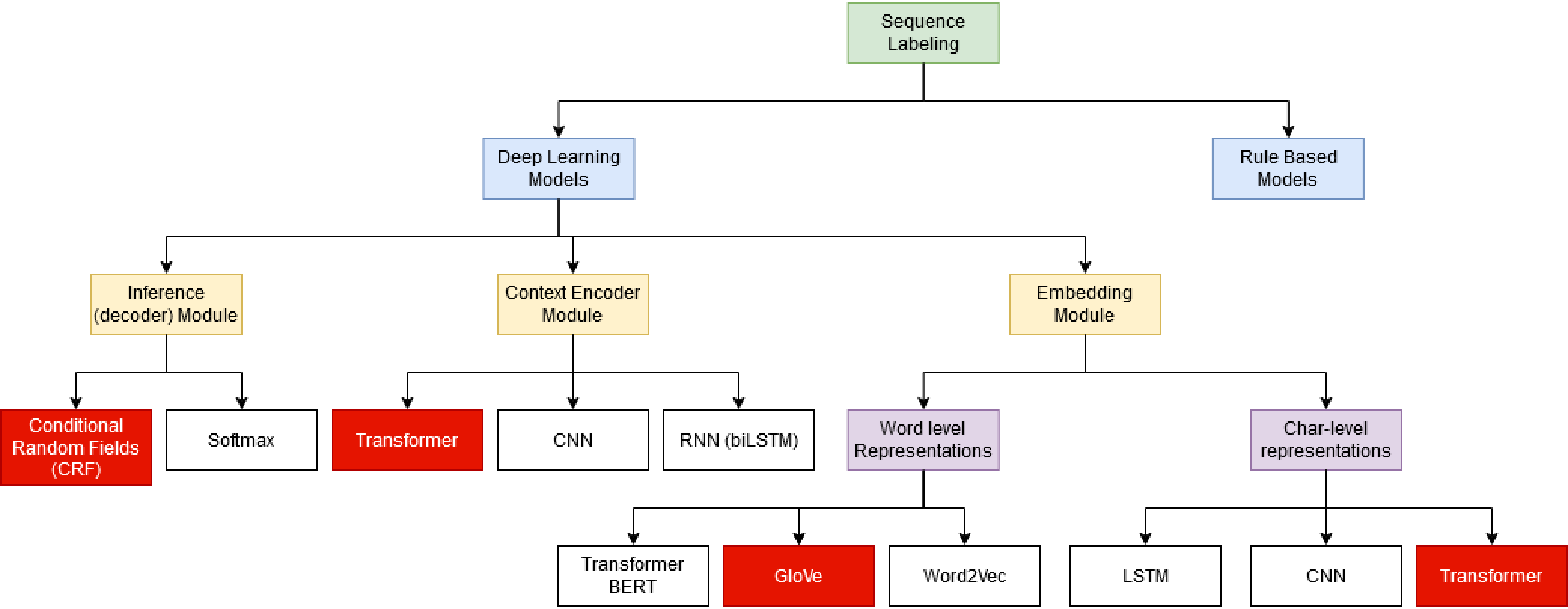
POS Tagging

External resources	Method	Accuracy
None	Collobert et al. 2011 [17]	97.29%
	Santos et al. 2014 [99]	97.32%
	Huang et al. 2015 [35]	97.55%
	Ling et al. 2015 [62]	97.78%
	Plank et al.2016 [88]	97.22%
	Rei et al. 2016 [92]	97.27%
	Vaswani et al. 2016 [110]	97.40%
	Andor et al. 2016 [2]	97.44%
	Ma and Hovy 2016 [71]	97.55%
	Ma and Sun 2016 [70]	97.56%
	Rei 2017 [91]	97.43%
	Yang et al. 2017 [120]	97.55%
	Kazi and Thompson 2017 [42]	97.37%
	Bohnet et al. 2018 [8]	97.96%
	Yasunaga et al. 2018 [121]	97.55%
	Liu et al. 2018 [65]	97.53%
	Zhang et al. 2018 [127]	97.59%
	Xin et al. 2018 [116]	97.58%
	Zhang et al. 2018 [128]	97.55%
	Hu et al. 2019 [34]	97.52%
	Cui et al. 2019 [18]	97.65%
	Jiang et al. 2020 [39]	97.7%
Unlabeled Word Corpus	Akbik et al. 2018 [1]	97.85%
	Clark et al. 2018 [15]	97.7%

POS tagging on PTB portion of WSJ data

TENER: Adapting Transformer Encoder for NER

Motivation		Transformer's low performance on NER
Approach		Char Embedding (Transformer) Word Embeddings (GloVe) Transformer CRF
Data		CoNLL-2003
Results		NER F1 =91.45

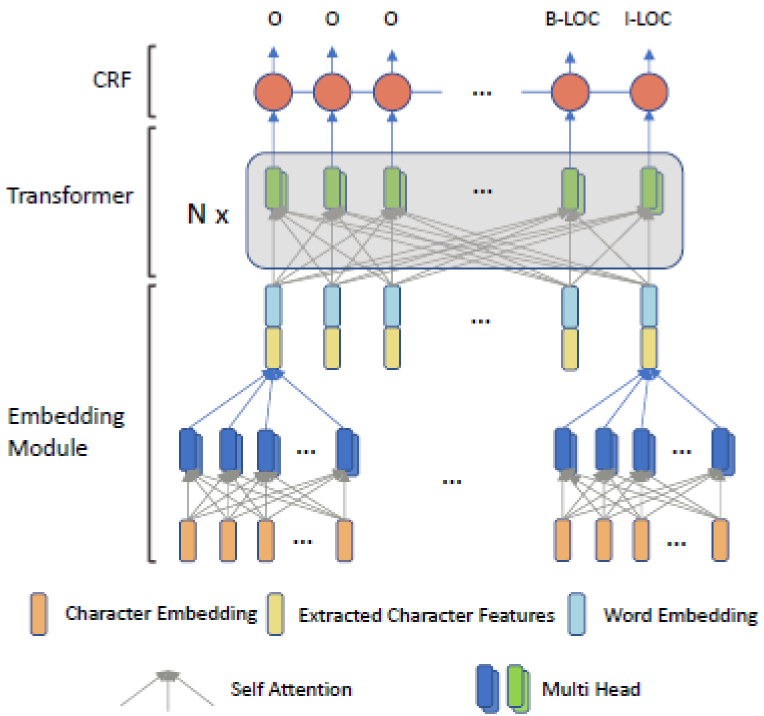


Sequence Labeling
Deep Learning Models
Rule Based Models
Inference (decoder) Module
Context Encoder Module
Embedding Module
Conditional Random Fields (CRF)
Softmax
Transformer
CNN
RNN (biLSTM)
Word level Representations
Char-level representations
Transformer BERT
GloVe
Word2Vec
LSTM
CNN
Transformer

Why transformers perform poorly on NER?

- 1. unaware of directionality
- 2. self attention is not aware of positions of different tokens
- 3. attention distribution is smooth and scaled
    - for NER sparse attention is suitable since all word not need to be attended
- Solution
    - abandon scale factor
    - use unscaled sharp attention

Architecture



Architecture

- Embedding Layer
  - CNN more efficient than BiLSTM
- Encoding Layer with Adapted Transformer
  - direction and distance aware
    - BiLSTM uses both sides
    - Transformer cannot distinguish which side the context information comes from
  - therefore, they changed the model
  - unscaled dot product attention
    - removed the scaling factor
    - sharper attention
    - beneficial only few words are named entities
- CRF

Results

Models	CoNLL2003	OntoNotes 5.0
BiLSTM-CRF (Huang et al., 2015)	88.83	
CNN-BiLSTM-CRF (Chiu and Nichols, 2016)	90.91 ± 0.20	86.12 ± 0.22
BiLSTM-BiLSTM-CRF (Lample et al., 2016)	90.94	
CNN-BiLSTM-CRF (Ma and Hovy, 2016)	91.21	
ID-CNN (Strubell et al., 2017)	90.54 ± 0.18	86.84 ± 0.19
LM-LSTM-CRF (Liu et al., 2018)	91.24 ± 0.12	
CRF+HSCRF (Ye and Ling, 2018)	91.26 ± 0.1	
BiLSTM-BiLSTM-CRF (Akhundov et al., 2018)	91.11	
LS+BiLSTM-CRF (Ghaddar and Langlais, 2018)	90.52 ± 0.20	86.57 ± 0.1
CN^3 (Liu et al., 2019)	91.1	
GRN (Chen et al., 2019)	91.44 ± 0.16	87.67 ± 0.17
Transformer	89.57 ± 0.12	86.73 ± 0.07
TENER (Ours)	91.33 ± 0.05	$\mathbf{88.43 \pm 0.12}$
w/ scale	91.06 ± 0.09	87.94 ± 0.1
w/ CNN-char	$\mathbf{91.45 \pm 0.07}$	88.25 ± 0.11

Different Word and Character level Embeddings

Char \ Word	BiLSTM	ID-CNN	AdaTrans
No Char	88.34 ± 0.32	87.30 ± 0.15	88.37 ± 0.27
BiLSTM	91.32 ± 0.13	89.99 ± 0.14	91.29 ± 0.12
CNN	91.22 ± 0.10	90.17 ± 0.02	$\mathbf{91.45 \pm 0.07}$
Transformer	91.12 ± 0.10	90.05 ± 0.13	91.23 ± 0.06
AdaTrans	91.38 ± 0.15	89.99 ± 0.05	91.33 ± 0.05

(a) CoNLL2003

- Problem with CNN for Char Embedding
    - cannot solve patters with uncontinious patters
        - **un….ily**
        - unhappily unnecessarily
- Transformer captures these patterns

8.Scientific BERT SCIBERT

Motivation

Pretraining language model for scientific texts

Using BERT

Approach

Word Embeddings (BERT)

BiLSTM

CRF

Data

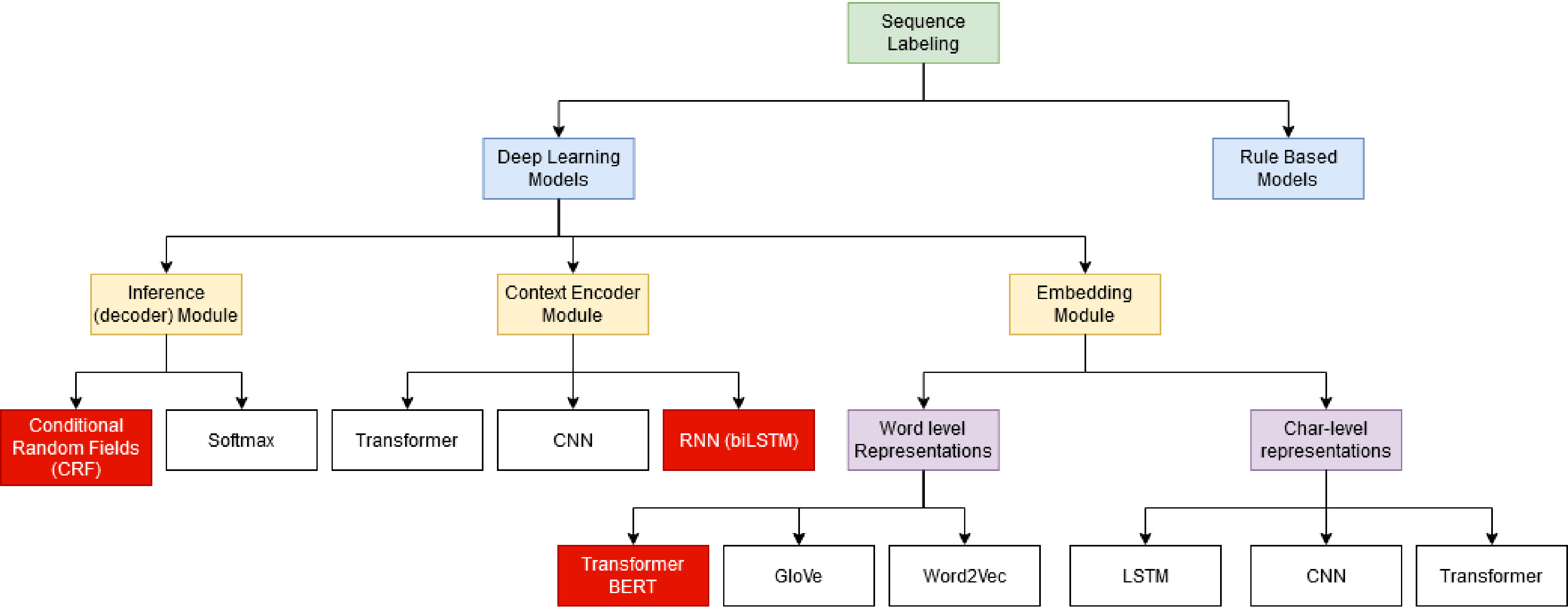
SciERC (CS domain)

EBM-NLP (biomedical domain)

Results

Biomedical : 1.1% increase in F1 score with respect to SOA

CS:2.5% increase in F1 score with respect to SOA



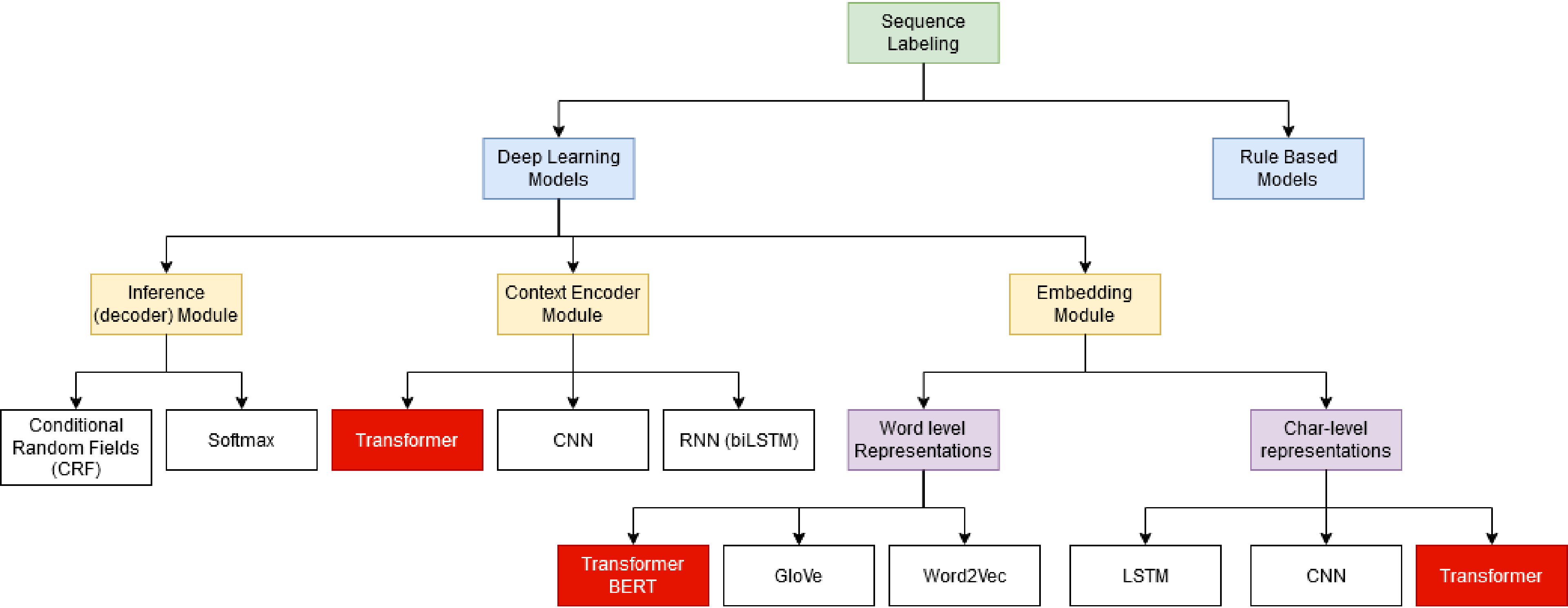
Sequence Labeling
Deep Learning Models
Rule Based Models
Inference (decoder) Module
Context Encoder Module
Embedding Module
Conditional Random Fields (CRF)
Softmax
Transformer
CNN
RNN (biLSTM)
Word level Representations
Char-level representations
Transformer BERT
GloVe
Word2Vec
LSTM
CNN
Transformer

Results

Field	Task	Dataset	SOTA	BERT-Base		SciBERT	
				Frozen	Finetune	Frozen	Finetune
Bio	NER	BC5CDR (Li et al., 2016)	88.85[7]	85.08	86.72	88.73	**90.01**
		JNLPBA (Collier and Kim, 2004)	**78.58**	74.05	76.09	75.77	77.28
		NCBI-disease (Dogan et al., 2014)	**89.36**	84.06	86.88	86.39	88.57
	PICO	EBM-NLP (Nye et al., 2018)	66.30	61.44	71.53	68.30	**72.28**
	DEP	GENIA (Kim et al., 2003) - LAS	**91.92**	90.22	90.33	90.36	90.43
		GENIA (Kim et al., 2003) - UAS	**92.84**	91.84	91.89	92.00	91.99
	REL	ChemProt (Kringelum et al., 2016)	76.68	68.21	79.14	75.03	**83.64**
CS	NER	SciERC (Luan et al., 2018)	64.20	63.58	65.24	65.77	**67.57**
	REL	SciERC (Luan et al., 2018)	n/a	72.74	78.71	75.25	**79.97**
	CLS	ACL-ARC (Jurgens et al., 2018)	67.9	62.04	63.91	60.74	**70.98**
Multi	CLS	Paper Field	n/a	63.64	65.37	64.38	**65.71**
		SciCite (Cohan et al., 2019)	84.0	84.31	84.85	**85.42**	**85.49**
Average				73.58	77.16	76.01	79.27

Small and Practical BERT Models for Sequence Labeling

Motivation	Faster smaller sequence labeling models for multilingual datasets
Approach	3 Layer BERT
Data	CoNLL-2018
Results	POS Accuracy = 94.5


Sequence Labeling
Deep Learning Models
Rule Based Models
Inference (decoder) Module
Context Encoder Module
Embedding Module
Conditional Random Fields (CRF)
Softmax
Transformer
CNN
RNN (biLSTM)
Word level Representations
Char-level representations
Transformer BERT
GloVe
Word2Vec
LSTM
CNN
Transformer

Results

Model	Multilingual?	Part-of-Speech F1	Morphology F1
Meta-LSTM	No	94.5	92.5
BERT	No	**95.1**	**93.0**
Meta-LSTM	Yes	91.1	82.9
BERT	Yes	**94.5**	**91.0**

References

1.How toextractunitof measurein scientificdocuments?ProceedingsoftheInternationalCon-ferenceon Knowledge DiscoveryandInformation RetrievalandtheInternational Conference on Knowl-edgeManagement andInformation Sharing, 2013.

2.S. Bozkurt, E. Alkim, I. Banerjee, andD. L. Rubin. Automateddetectionof measurementsandtheirdescriptorsin radiologyreportsusinga hybridnaturallanguageprocessingalgo-rithm.Journalof DigitalImaging, 32(4):544–553, 2019.

3.G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, andC. Dyer. Neuralarchitecturesfornamedentityrecognition, 2016.

4.X. MaandE. Hovy. End-to-endsequencelabelingviabi-directionallstm-cnns-crf, 2016.

5.Guo, Qipeng& Qiu, Xipeng& Liu, Pengfei& Shao, Yunfan& Xue, Xiangyang& Zhang, Zheng. (2019). Star-Transformer.

6.He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., & Jiang, S. (2020, November 13). A Survey on Recent Advances in Sequence Labeling from Deep Learning Models. Retrieved December 06, 2020, from

7.Yan, Hang & Deng, Bocao& Li, Xiaonan& Qiu, Xipeng. (2019). TENER: Adapting Transformer Encoder for Name Entity Recognition.

8.Beltagy, Iz& Cohan, Arman & Lo, Kyle. (2019). SciBERT: Pretrained Contextualized Embeddings for Scientific Text.

9.Tsai, Henry & Riesa, Jason & Johnson, Melvin & Arivazhagan, Naveen & Li, Xin & Archer, Amelia. (2019). Small and Practical BERT Models for Sequence Labeling.