Natural Language Processing with Deep Learning

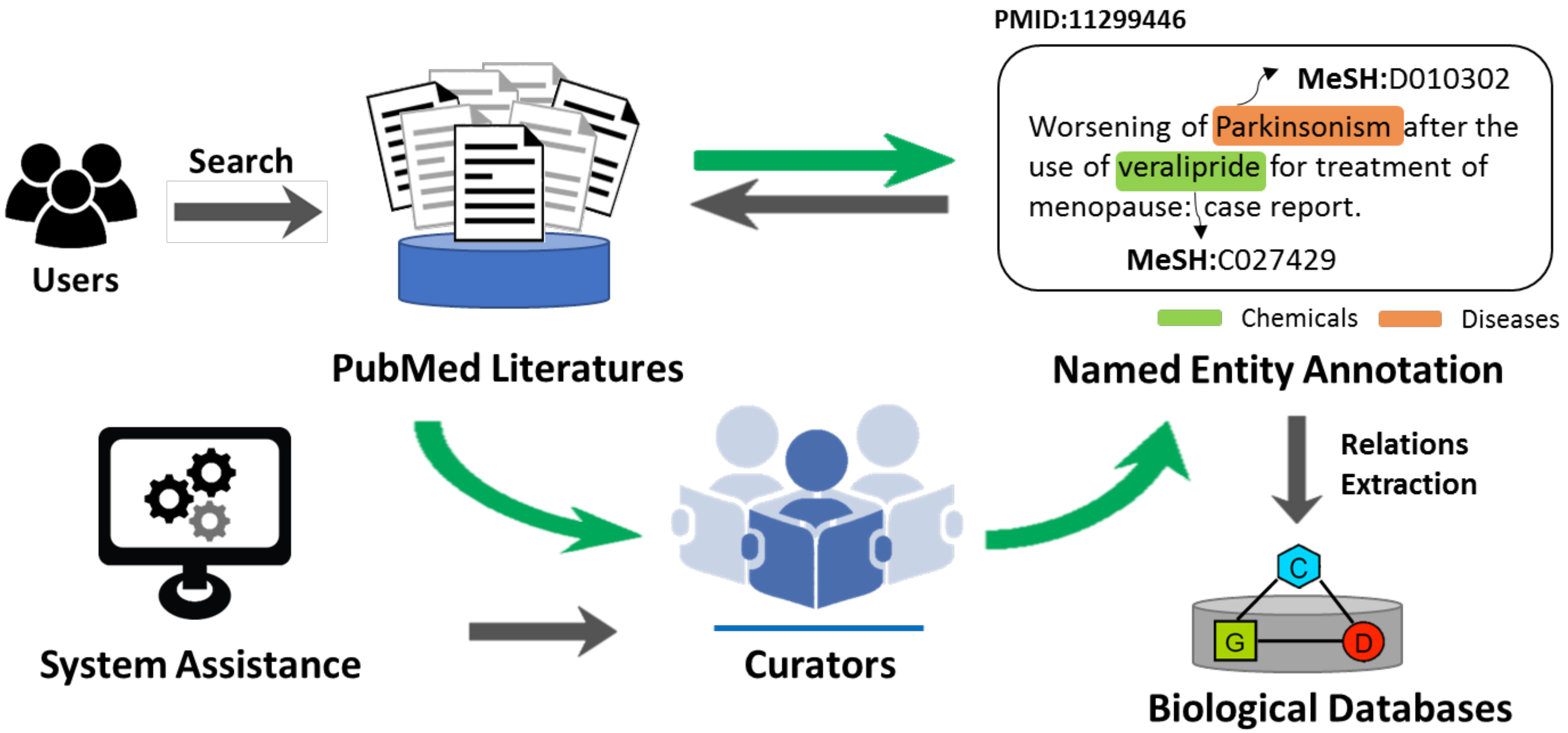
Hung-Yu Kao (高宏宇)
***I**ntelligent **K**nowledge **M**anagement Lab*

Institute of Medical Informatics,
Dept. of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan

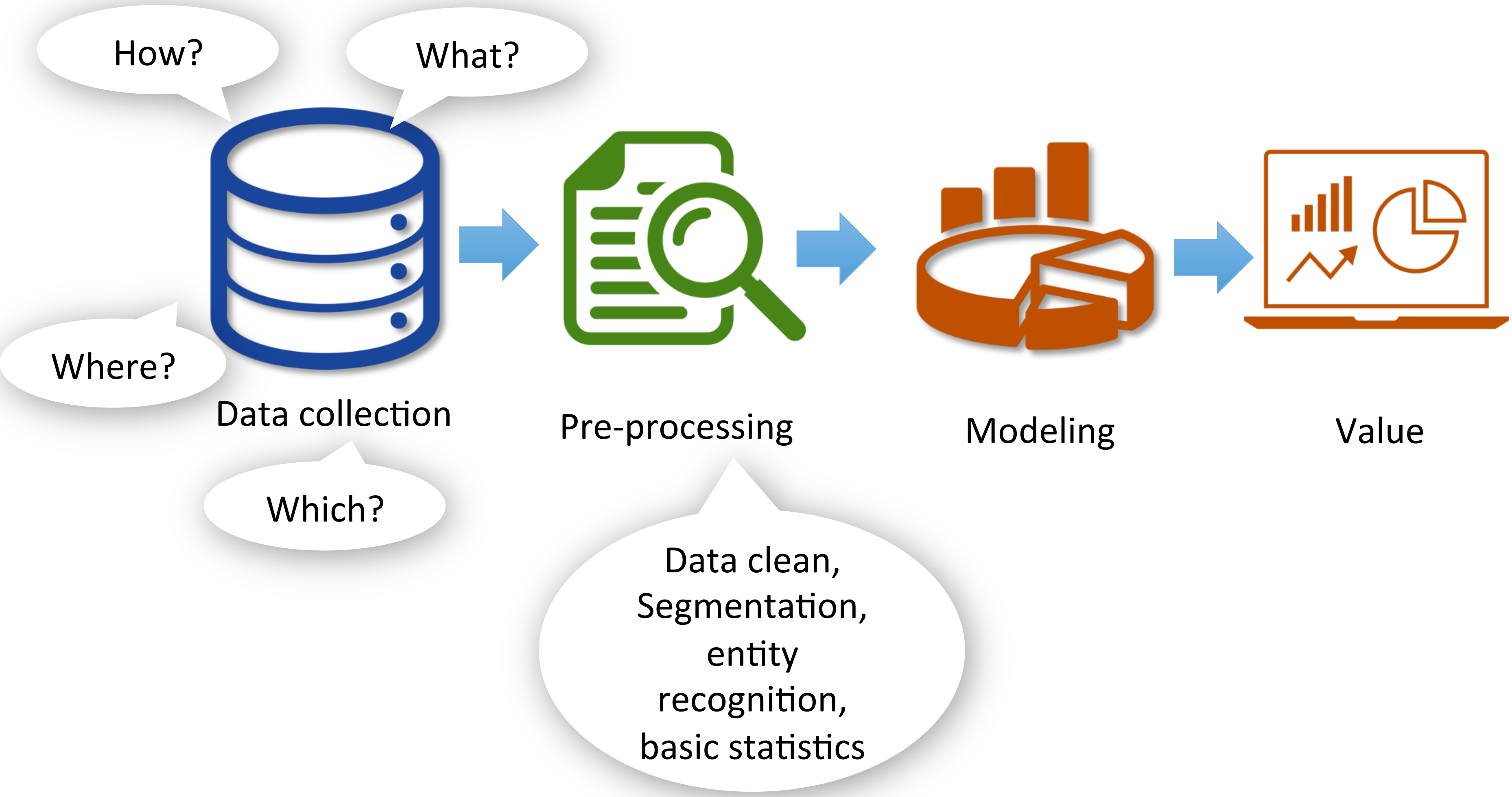
Outlines

- NLP concepts and issues
- Word Representation
- Sentence Representation
- Applications
  - Named Entity Recognition
  - Rumor Detection
  - Chatbot Training by GAN

Biomedical Text mining process / issues

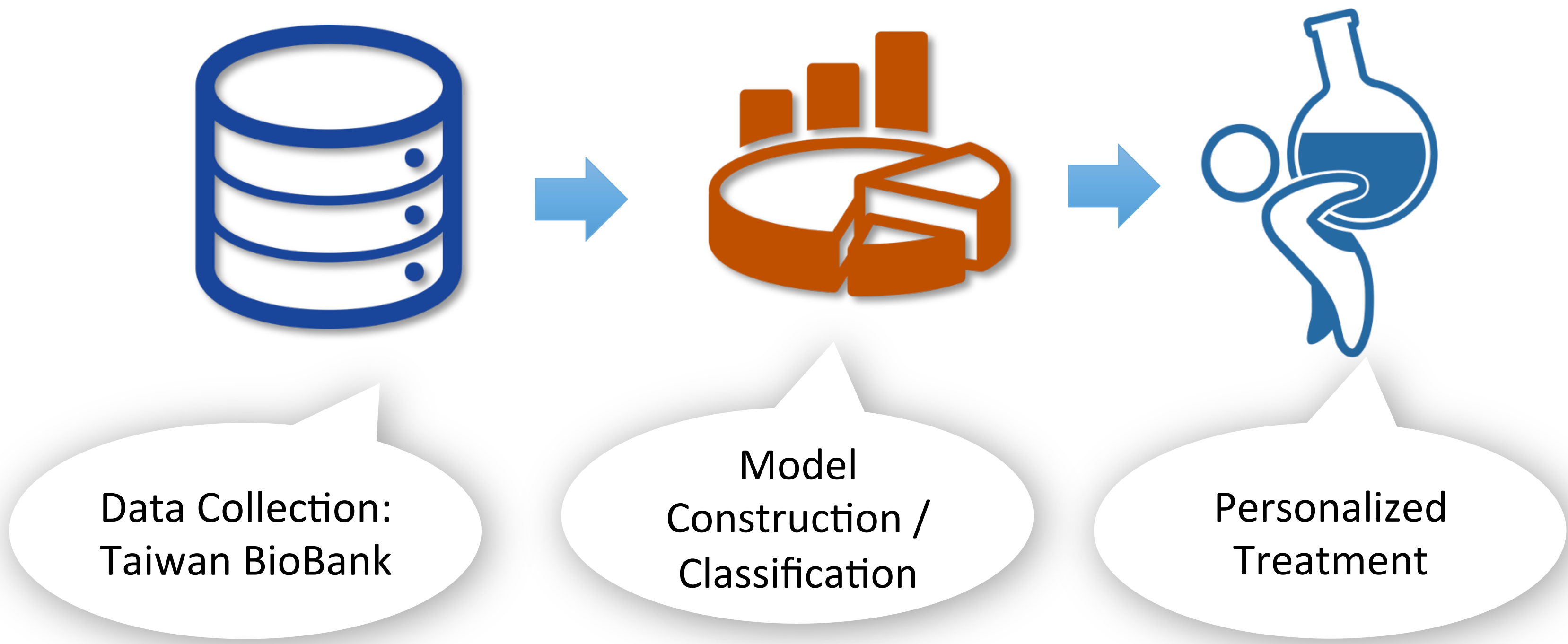


Text mining process / issues



CRF-based named entity recognition

Precision Medicine



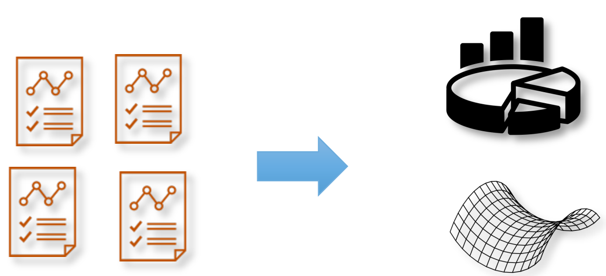
NIPS 2017 competition

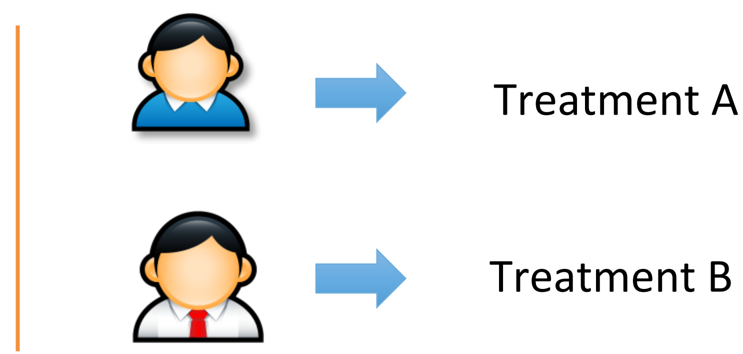
-- Personalized Medicine: Redefining Cancer Treatment

- **Launched by Memorial Sloan Kettering Cancer Center (MSKCC)**

- Once sequenced, a cancer tumor can have thousands of genetic mutations. But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers).
    - This is a very time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature.

- **Prize $15,000**





NIPS 2017 competition

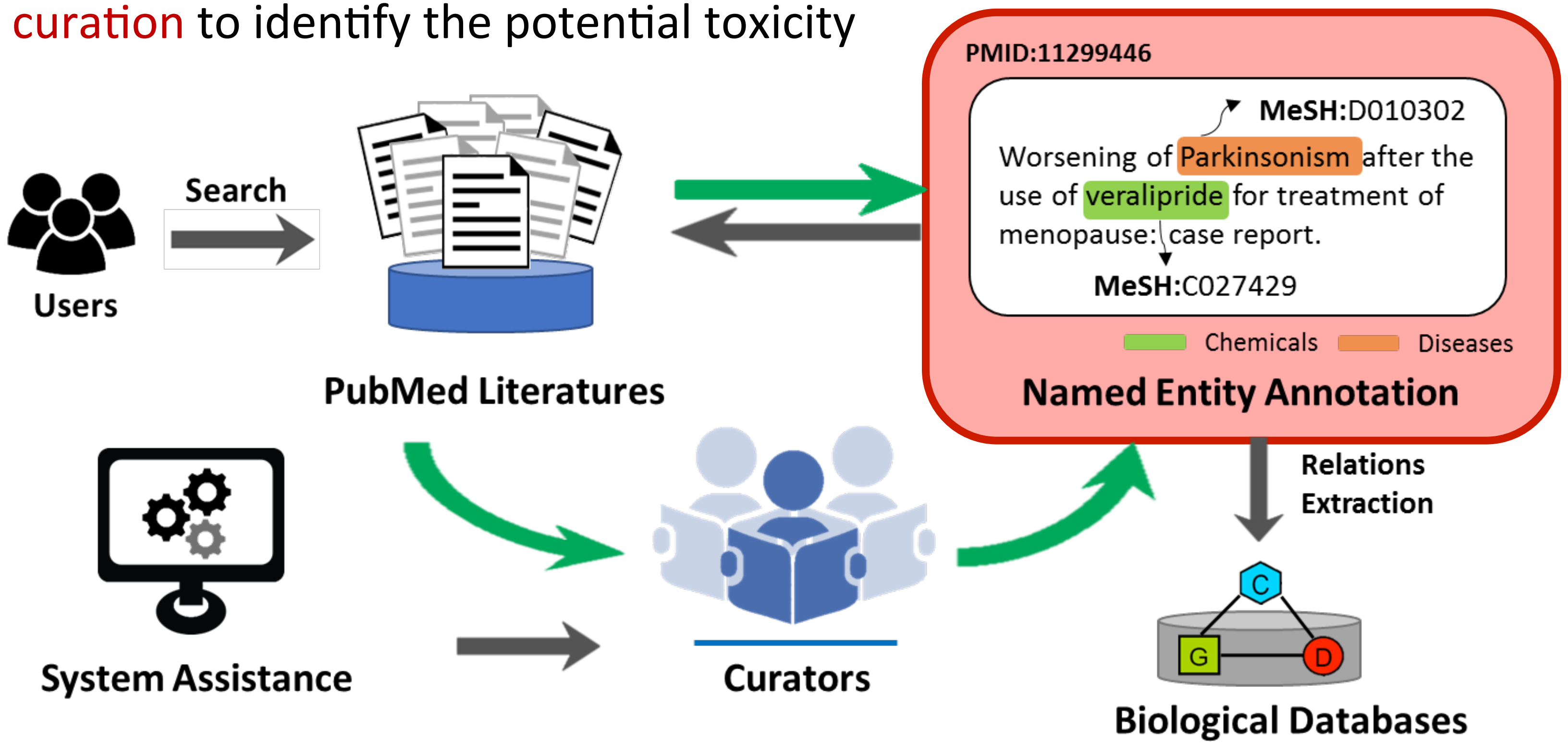
-- Personalized Medicine: Redefining Cancer Treatment

0||Cyclin-dependent kinases (CDKs) regulate a variety of fundamental cellular processes. CDK10 stands out as one of the last orphan CDKs for which no activating cyclin has been identified and no kinase activity revealed. Previous work has shown that CDK10 silencing increases ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2)-driven activation of the MAPK pathway, which confers tamoxifen resistance to breast cancer cells. The precise mechanisms by which CDK10 modulates ETS2 activity, and more generally the functions of CDK10, remain elusive. Here we demonstrate that CDK10 is a cyclin-dependent kinase by identifying cyclin M as an activating cyclin. Cyclin M, an orphan cyclin, is the product of FAM58A, whose mutations cause STAR syndrome, a human developmental anomaly whose features include toe syndactyly, telecanthus, and anogenital and renal malformations. We show that STAR syndrome-associated cyclin M mutants are unable to interact with CDK10. Cyclin M silencing phenocopies CDK10 silencing in increasing c-Raf and in conferring tamoxifen resistance to breast cancer cells. CDK10/cyclin M phosphorylates ETS2 in vitro, and in cells it positively controls ETS2 degradation by the proteasome. ETS2 protein levels are increased in cells derived from a STAR patient, and this increase is

Introduction

- **The importance of text mining in biomedical field:**

  Mining useful knowledge from the biomedical literature holds potentials for facilitating literature search and biological database curation to identify the potential toxicity



Introduction

- **The problem of biocuration:**

  A rapid literature growth in the biomedical field causes problems on aggregating knowledge to biocurators.

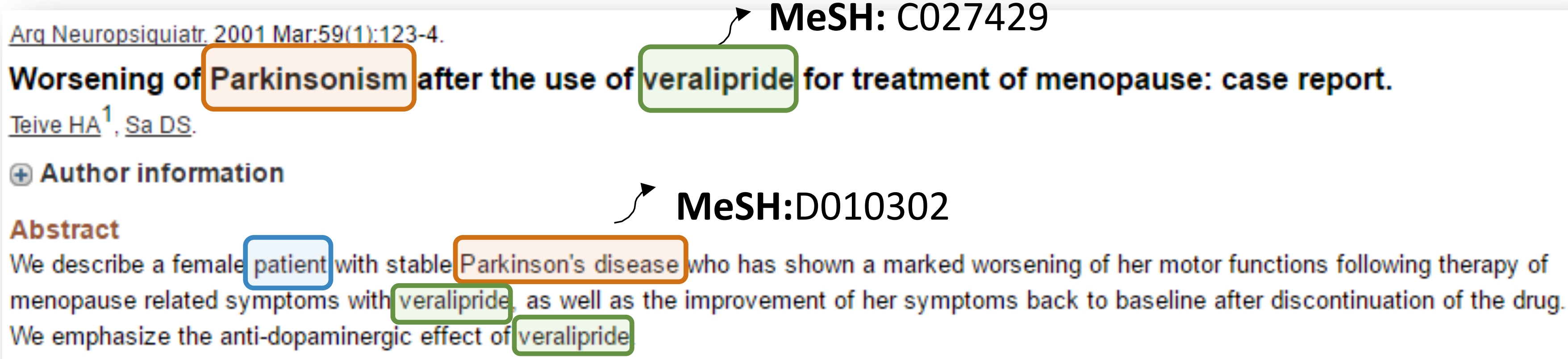
  - Time-consuming
  - Expensive

- **The example of biocuration from PubMed Abstract:**
  - PMID:11299446

Chemicals Diseases Species

MeSH: C027429

MeSH:D010302



Arq Neuropsiquiatr. 2001 Mar;59(1):123-4.

Worsening of Parkinsonism after the use of veralipride for treatment of menopause: case report.

Teive HA[1], Sa DS.

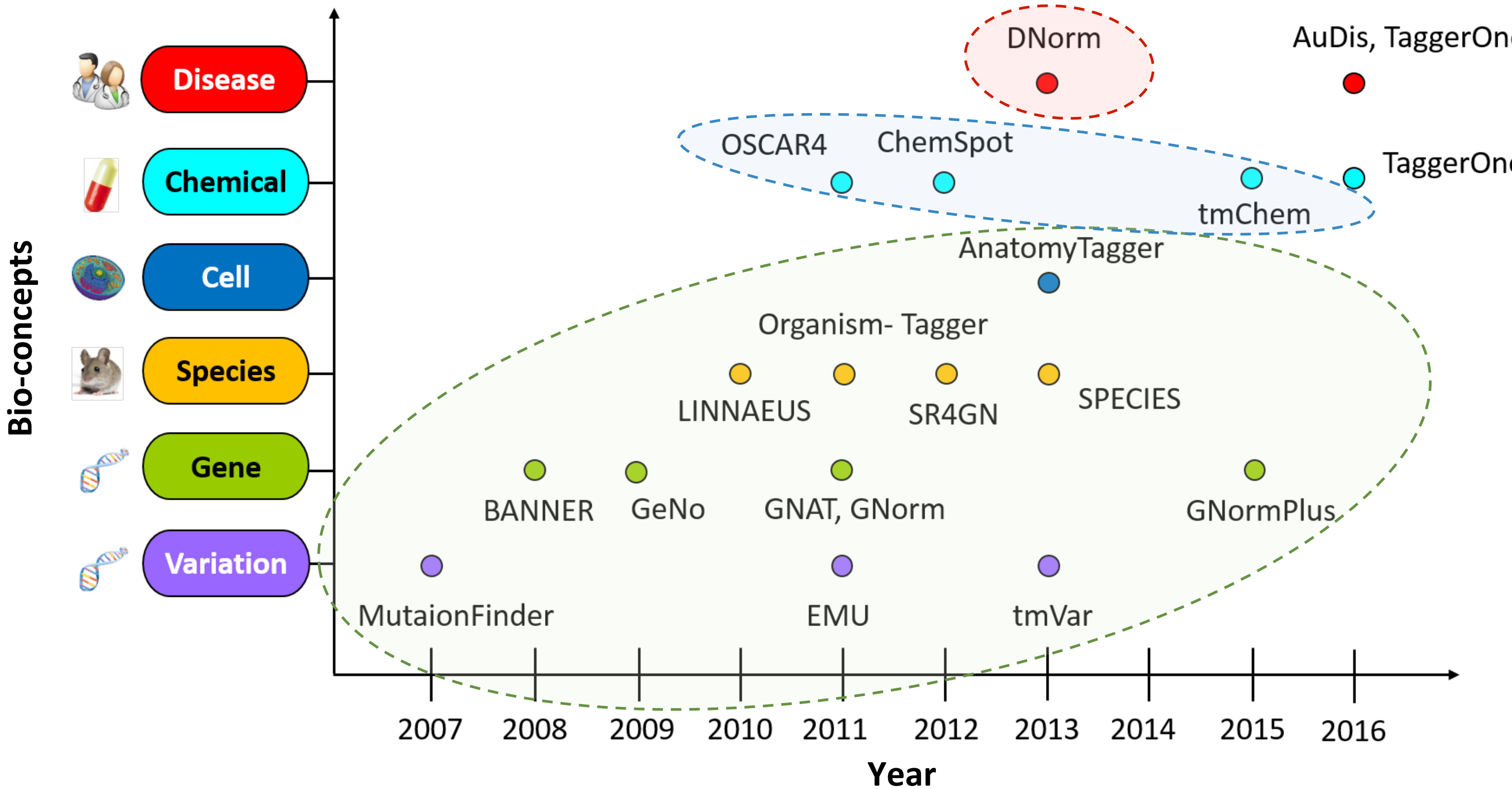
Author information

Abstract

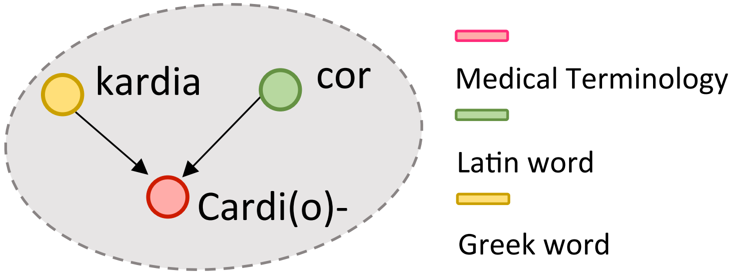
We describe a female patient with stable Parkinson's disease who has shown a marked worsening of her motor functions following therapy of menopause related symptoms with veralipride, as well as the improvement of her symptoms back to baseline after discontinuation of the drug. We emphasize the anti-dopaminergic effect of veralipride.

Motivation

- The named-entity recognition (NER) task is a prerequisite for the analysis of high-throughput screen and cross-referencing in databases
- Most systems are specialized to a single bio-concept

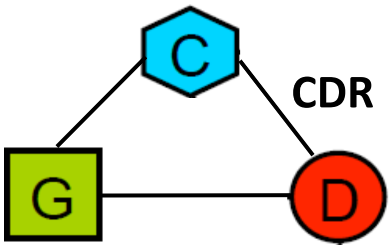


Motivation - *Difficulties*



- Before extracting CID relation, recognition of disease and chemical entities is needed.
- **Challenges of extracting disease named entities:**
  - Medical terminology should be collected and organized by the schemes of medical nomenclature.
    - Human body (e.g., 'cardi(o)-' refers to 'heart')
    - Disease
    - Symptom
    - Specific terminologies (e.g., 'Code Blue' refers to 'cardiopulmonary arrest')
  - Combined by medical roots, suffixes and prefixes derived from ancient Greek or classical Latin.
    - Diversity and difficult to be effectively recognized by machine learning.

Motivation - *Difficulties*



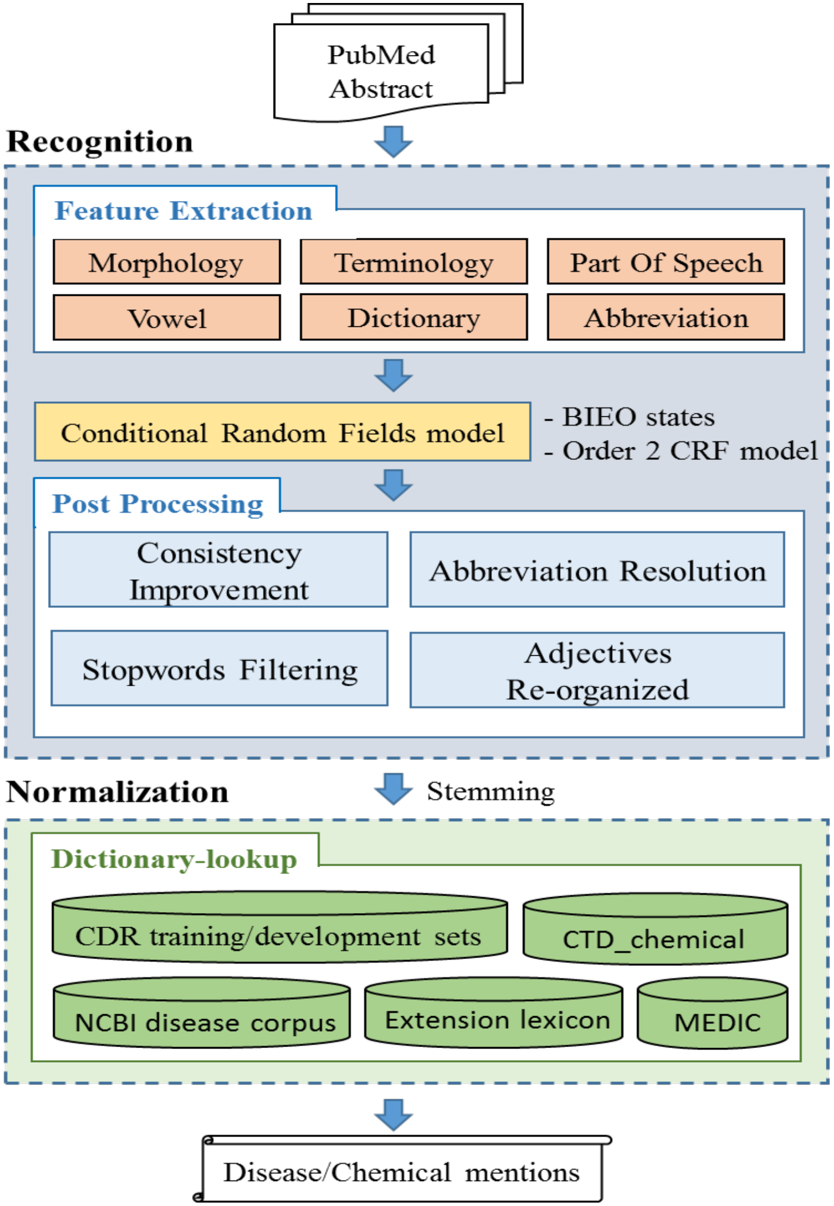
- Lack of Disease Corpus
- A great diversity of disease names:
  1. **Disease terminology:**
     - E.g. ‘cancer’, ‘carcinoma’, and ‘malignant tumour’
  2. **Combination word:**
     - E.g. ‘nephro’ → ‘nephritis’ and ‘nephropathy’
  3. **Abbreviation : Diseases ? Genes ?**
     - E.g. ‘HD’ → ’Huntington disease’
  4. **Composite disease mention**
     - E.g. ‘ovarian and peritoneal cancer’ → ‘ovarian cancer’ and ‘peritoneal cancer’
  5. **Different writing order:**
     - E.g. ‘carcinoma of the breast’ = ‘Breast carcinoma’

Nephr(o)-	-itis	-pathy
Kidney	inflammation	disease

- **Assign the mention a relative database identifier (MeSH)**
  - Facilitate the extraction of Chemical-induced Disease Relation (CDR)

Methods

- **Recognition**:
  machine learning-based
  - 6 feature groups
  - CRF model
  - Post Processing
- **Normalization**:
  - Lexicon extension
  - Dictionary-lookup



CRF Module

- **General tokenization**
  - White spaces
  - Punctuations
  - Digits

E.g. Amylo-1,6-Glucosidase Deficiency

Amylo	-	1	,	6	-	Glucosidase	Deficiency

Title	Risk factors and predictors of levodopa-induced dyskinesia among multiethnic Malaysians with Parkinson's disease.
Abstract	Chronic pulsatile levodopa therapy for Parkinson's disease (PD) leads to the development of motor fluctuations and dyskinesia……..

PMID	**START OFFSET**	**END OFFSET**	**MENTION**	**MENTION TYPE**	**DATABASE IDENTIFIER**
23952588	31	39	levodopa	Chemical	D007980
23952588	48	58	dyskinesia	Disease	D004409
23952588	93	112	Parkinson's disease	Disease	D010300
23952588	132	140	levodopa	Chemical	D007980

- **CRF states**
  - BIEO states (B: begin, I: insides, E: end , O: outside)
  - Tag chemical/disease type with BIE states (e.g. B_Chem)

Predictors of levodopa-induced dyskinesia among multiethnic Malaysians with Parkinson's disease

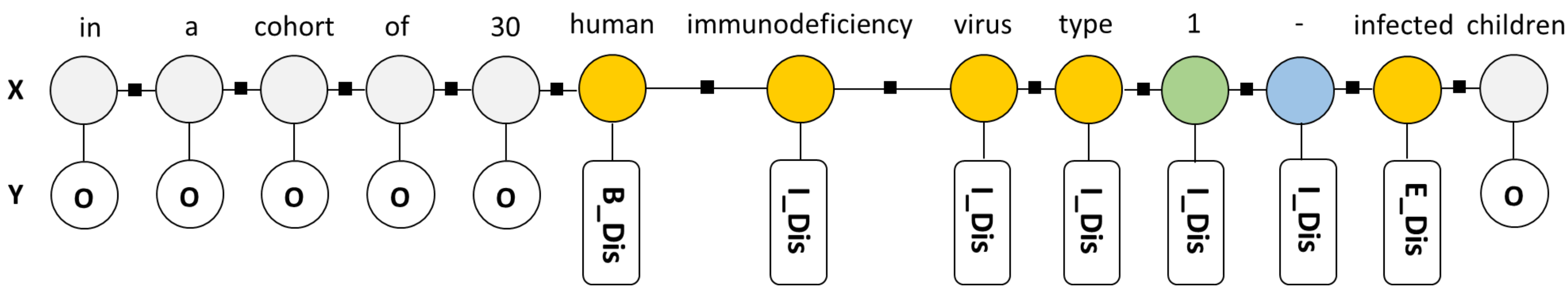
Predictors	of	levodopa	-	induced	dyskinesia	among	multiethnic	Malaysians	with	Parkinson	'	s	disease
O	O	B_Chem	O	O	B_Dis	O	O	O	O	B_Dis	I_Dis	I_Dis	E_Dis

CRF Module

- **Conditional random fields (CRFs)**

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\exp(f(\boldsymbol{Y},\boldsymbol{X}))}{\sum_{Y'}\exp(f(\boldsymbol{Y'},\boldsymbol{X}))}$$

  - $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ is a token(observation) sequence.
  - $\mathbf{Y} = \{y_1, y_2, \dots, y_n\}$ is a label(state) sequence from X.



Lafferty, J., McCallum, A. and Pereira, F. 2001 Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML 01*

CRF Module

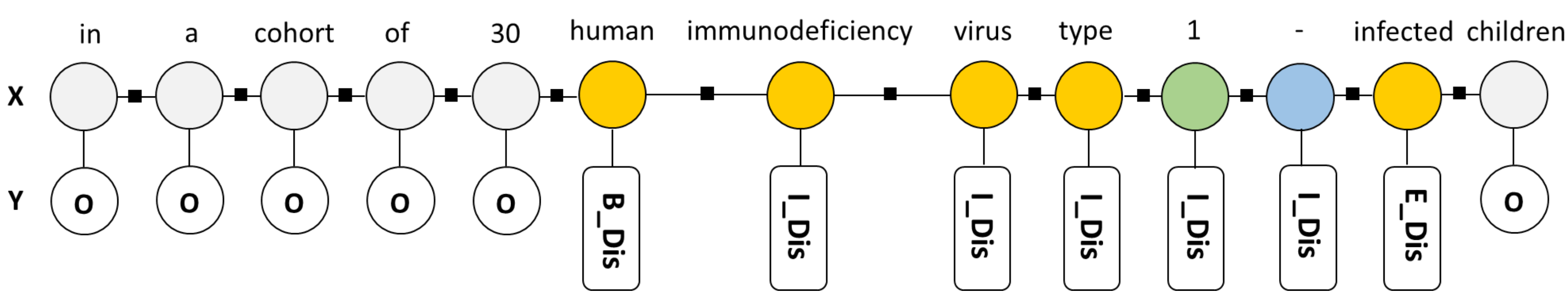
- **Conditional random fields (CRFs)**

- $f(\boldsymbol{Y},\boldsymbol{X}) = \sum_{j=1}^{n}\sum_{i=1}^{m}\omega_i f_i(y_j, y_{j-1}, \boldsymbol{X})$ is a weighted feature function consists of state and transition feature function between position j and $j-1$
- $\omega(\omega_1, \dots, \omega_l)$ is a feature weight vector.
- **Feature function:**

$$f_i(y_j, y_{j-1}, \mathbf{X}) = \begin{cases} D(\mathbf{X}, i, j) & y_j = s,\ y_{j-1} = t \quad s, t \in \text{STATES} \\ 0 & \text{otherwise} \end{cases}$$

- **Observation function:**

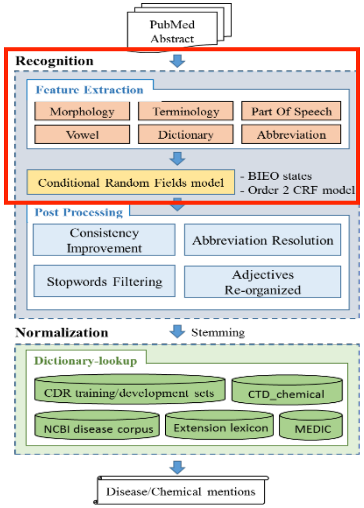
$$D(\mathbf{X}, i, j) = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ token in } \mathbf{X} \text{ match to feature } i \\ 0 & \text{otherwise} \end{cases}$$



Lafferty, J., McCallum, A. and Pereira, F. 2001 Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML 01*

Feature extraction -

6 feature groups in CRF model



- **<u>Morphology</u>:**
  - Original tokens
  - Stemmed tokens (*Snowball library*)
  - prefixes/suffixes (length 1 to 5)
- **<u>Part Of Speech:</u>** (*Stanford tagger*)
  - noun, verb, adjective etc

Feature extraction

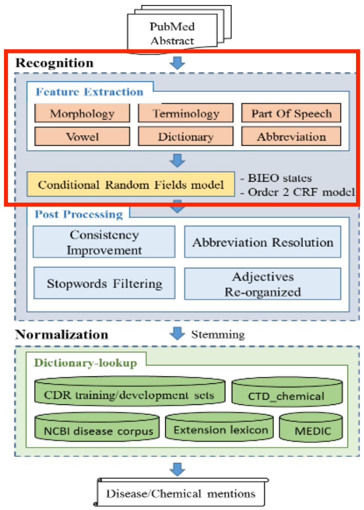
- The *6 feature groups* we utilized in CRF model:

Sequence

Morphology												Terminology	Part Of Speech	Vowel	Dictionary-lookup	Abbreviation	State
Scleroderma	Scleroderma	S	Sc	Scl	Scle	Scler	a	ma	rma	erma	derma	-N/A-	NNP	v:Scl_r_d_rm_	DIC:Y	Fullname	B
renal	renal	r	re	ren	rena	renal	l	al	nal	enal	renal	-body part-	JJ	v:r_n_l	DIC:Y	Fullname	I
crisis	crisi	c	cr	cri	cris	crisi	s	is	sis	isis	risis	-N/A-	NN	v:cr_s_s	DIC:Y	Fullname	E
(	(	(	N/A	N/A	N/A	N/A	(	N/A	N/A	N/A	N/A	-N/A-	(	v:(	N/A	N/A	O
SRC	SRC	S	SR	SRC	N/A	N/A	C	RC	SRC	N/A	N/A	-N/A-	NNP	v:SRC	DIC:Y	Abbreviation	B
)	)	)	N/A	N/A	N/A	N/A	)	N/A	N/A	N/A	N/A	-N/A-	)	v:)	N/A	N/A	O
is	is	i	is	N/A	N/A	N/A	s	is	N/A	N/A	N/A	-N/A-	VBZ	v:_s	N/A	N/A	O
a	a	a	N/A	N/A	N/A	N/A	a	N/A	N/A	N/A	N/A	-N/A-	DT	v:_	N/A	N/A	O
rare	rare	r	ra	rar	rare	N/A	e	re	are	rare	N/A	-N/A-	JJ	v:r_r_	N/A	N/A	O
complication	complic	c	co	com	comp	compl	n	on	ion	tion	ation	-N/A-	NN	v:c_mpl_c_t_n	N/A	N/A	O
of	of	o	of	N/A	N/A	N/A	f	of	N/A	N/A	N/A	-N/A-	IN	v:_f	N/A	N/A	O
systemic	systemic	s	sy	sys	syst	syste	c	ic	mic	emic	temic	-N/A-	JJ	v:syst_m_c	DIC:Y	Fullname	B
sclerosis	sclerosi	s	sc	scl	scle	scler	s	is	sis	osis	rosis	-N/A-	NN	v:scl_r_s_s	DIC:Y	Fullname	E
(	(	(	N/A	N/A	N/A	N/A	(	N/A	N/A	N/A	N/A	-N/A-	(	v:(	N/A	N/A	O
SSc	SSc	S	SS	SSc	N/A	N/A	c	Sc	SSc	N/A	N/A	-N/A-	NNP	v:SSc	DIC:Y	Abbreviation	B
)	)	)	N/A	N/A	N/A	N/A	)	N/A	N/A	N/A	N/A	-N/A-	)	v:)	N/A	N/A	O

Feature extraction -

6 feature groups in CRF model



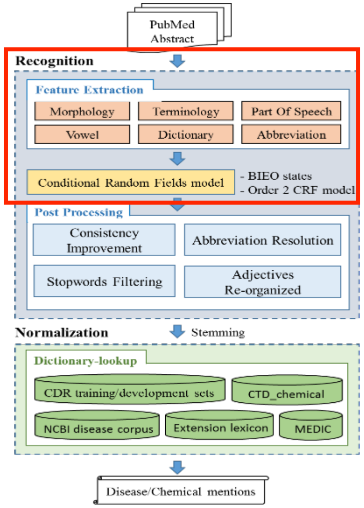
- **Terminology:**

Disease

Groups	Conditions
disease terminologies	impairment, nausea, vomiting, disease, cancer, toxicity, insufficiency, effusion, deficit, dysfunction, injury, pain, neurotoxicity, infect, syndrome, symptom, hyperplasia, retinoblastoma, defect, disorder, failure, hamartoma, hepatitis, tumor, damage, illness, abnormality, tumour, abortion
body part	pulmonary, neuronocular, orbital, breast, renal, hepatic, liver, hart, eye, pulmonary, ureter, bladder, pleural, pericardial, colorectal, head, neck, pancreaticobiliary, cardiac, leg, back, cardiovascular, gastrointestinal, myocardial, kidney, bile, intrahepatic, extrahepatic, memorygastric
human ability	visual, auditory, learning, opisthotonu, sensory, motor, memory, social, emotion

Feature extraction -

6 feature groups in CRF model



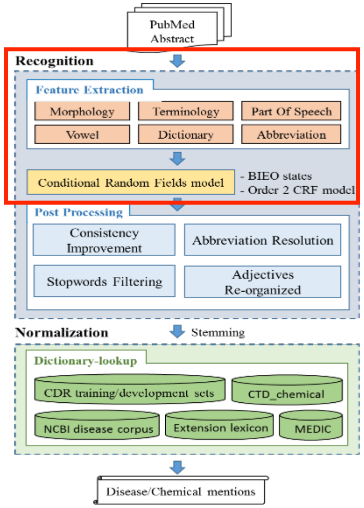
- **Terminology:**

Chemical

Groups	Condition
CHEM Inline Suffix	yl, ylidyne, oyl, sulfonyl
CHEM Alkane Stem	meth, eth, prop, tetracos
CHEM Simple Multiplier	di, tri, tetra
CHEM Trivial Ring	benzen, pyridin, toluen
CHEM Suffix	ol, carboxylic, amide, ate, acid, and etc.
CHEM Suffix2	vir, cillin, mab, olol, tidine, and etc.
CHEM Element	hydrogen, helium, lithium, beryllium, and etc.
CHEM Element2	ydrogen, elium, ithium, eryllium, and etc.

Feature extraction -

6 feature groups in CRF model

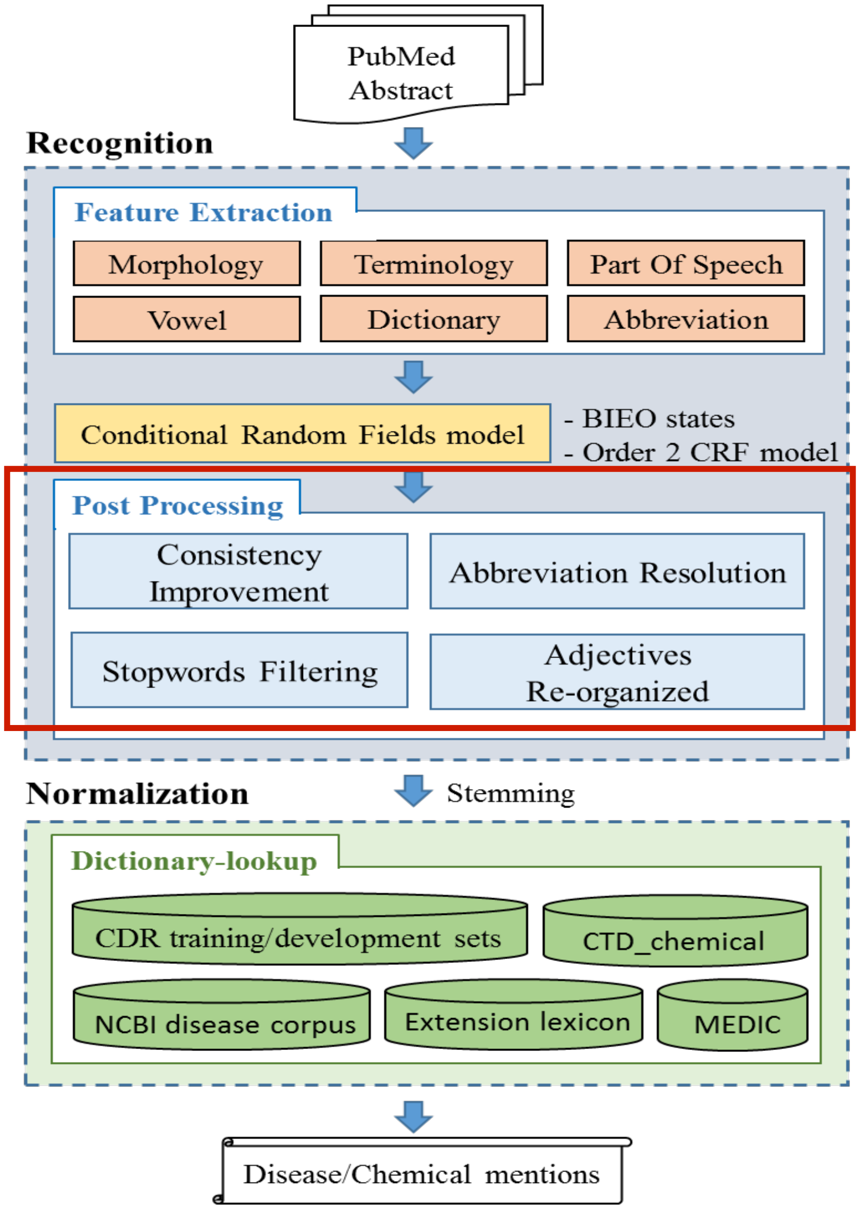


- **Vowel:**
  - vowels change to “-”
  - “tumor” and “tumour” are turned into “t-m-r“
- **Dictionary-lookup:**
  - CTD disease vocabulary (MEDIC) , CTD chemical vocabulary , NCBI disease corpus
  - The length greater than 3
- **Abbreviation detection:** (*BIOADI*)
  - Abbreviations
  - Full names

C.-J. Kuo, M. H. Ling, K.-T. Lin, and C.-N. Hsu, "BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature," BMC bioinformatics, vol. 10, p. S7, 2009.

Post-Processing

- Consistency improvement
- Abbreviation solution
- Adjectives Re-organized
- Stopwords filtering

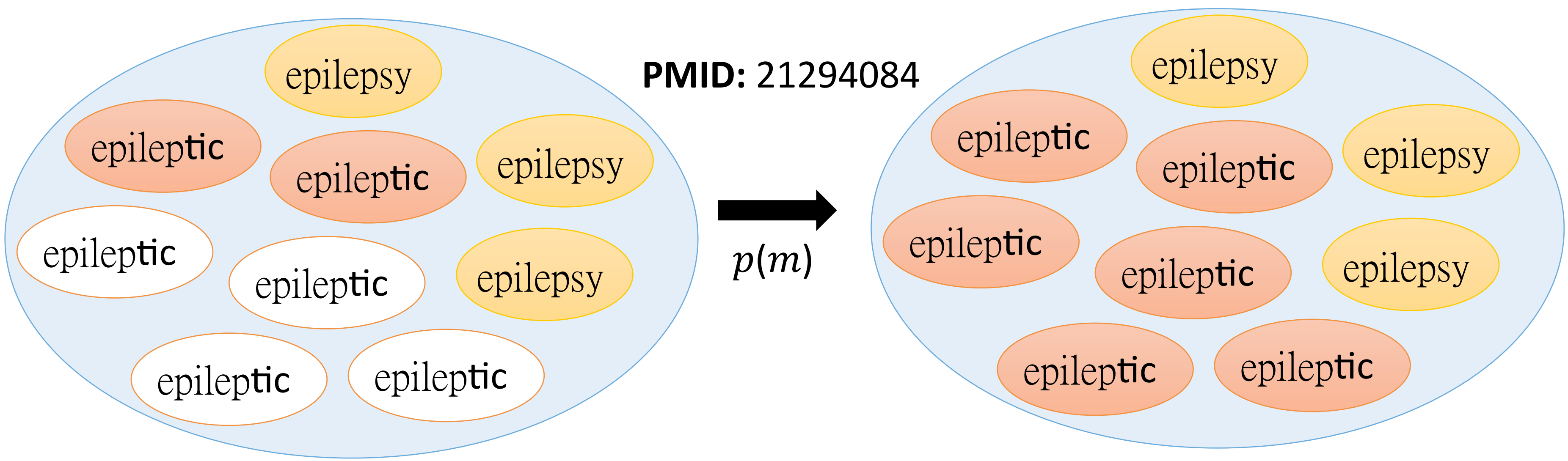


- Training data : CDR TrainingSet
- Testing data : CDR DevelopmentSet

Post-Processing

- **Consistency improvement :**
  - Detect the numbers of each mention in an abstract

$p(m)$= *quantity of CRF tagging/actual quantity of each mention*
$> \alpha$ *, where* α=0.25



Disease mentions detected by CRF

Consistency improvement

Post-Processing

- **Abbreviation resolution**

1) The long form (LF) of the pair is recognized as a disease mention
2) The short form (SF) is recognized as a disease mention
3) Check whether the LF contains disease terminologies or not

Type 1:

Long Form	Short Form
subarachnoid hemorrhage	(SAH)

Type 2:

Long Form	Short Form
subarachnoid hemorrhage	(SAH)

Type 3:

Long Form	SF
Parkinson's Disease	(PD)

Terminology detection

Post-Processing

hepatic	tumors
O	**B_Disease**

↓

hepatic	tumors
B_Disease	**E_Disease**

- **Adjectives Re-organized:**
  - Part Of Speech feature in CRF
  - Terminology list in CRF

Adjective	Conjunction	Adjective	Nounce	Con.	Nounce	Con.	Nounce

- hepatic tumors
- Necrotising fasciitis
- vasogenic edema

- sensory and motor dysfunction
- glomerular or tubular dysfunction
- schizo-affective disorder

- ischemia of the globus pallidus

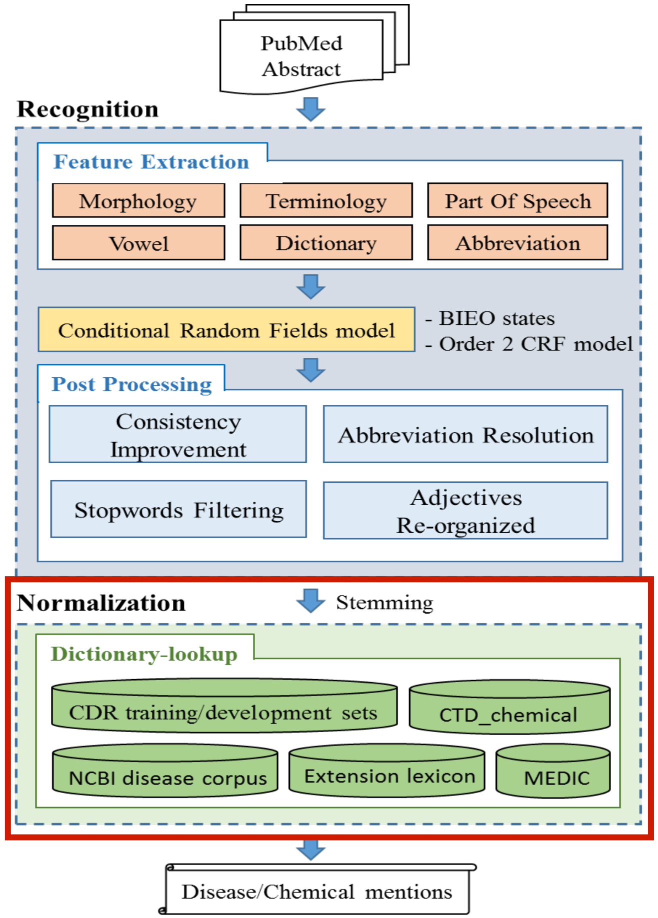
- tuberculosis of the lung or lymph node

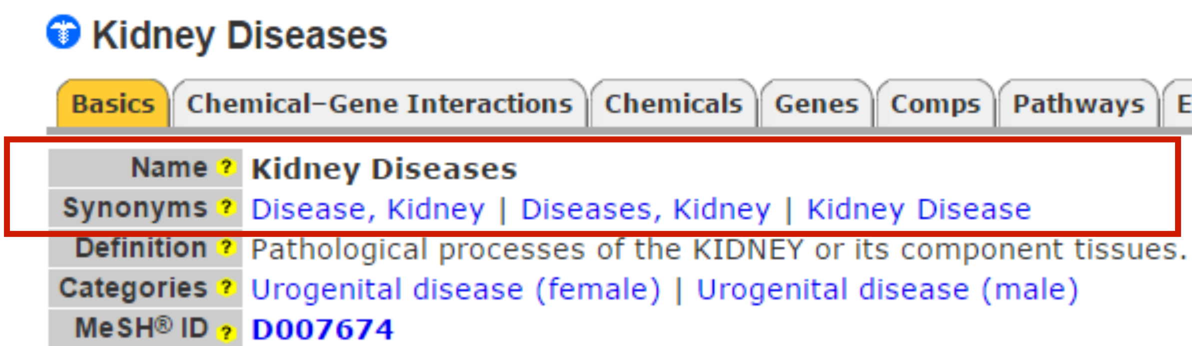
Normalization

- **Nomenclature-based Similarity**

Lexicon:

- MEDIC (Comparative Toxicogenomics Database)
- NCBI disease corpus
- CDR training/development sets
- MEDIC Synonyms
- MEDIC Extension
- CTD Chemical vocabulary





Experiment – *Datasets and Models*

- **BioCreative V – CDR Corpus**

Task Dataset	Articles	Chemical		Disease		CID
		Mention	ID	Mention	ID	relation
Training	500	5,203	1,467	4,182	1,965	1,038
Development	500	5,347	1,507	4,244	1,865	1,012
Test	500	5,385	1,435	4,424	1,988	1,066
	1,000	15,935	4,409	12,850	5,818	3,116

- **3 Models:**

Chemical-Disease **Disease** **Chemical**

Introduction

- CDR Corpus:
  - PMID|t|Title
  - PMID|a|Text
  - PMID start_offset end_offset token_name type MeSH

```
354896|t|Lidocaine-induced cardiac asystole.
354896|a|Intravenous administration of a single 50-mg bolus of lid
of the sinoatrial and atrioventricular nodal pacemakers. The patie
to the development of bradyarrhythmias; and, thus, this probably r
354896  0    9    Lidocaine    Chemical    D008012
354896  18   34   cardiac asystole    Disease D006323
354896  90   99   lidocaine    Chemical    D008012
354896  142  152  depression   Disease D003866
354896  331  347  bradyarrhythmias    Disease D001919
354896  409  418  lidocaine    Chemical    D008012
```

Evaluation Results – *Model 1*

- Chemical-Disease Model
- Evaluation on disease:

*The highest value is shown in bold

	Tools	Precision	Recall	F-score
NER	CDRnN	**0.869**	**0.831**	**0.850**
	TaggerOne	0.847	0.810	0.828
NEN	CDRnN	**0.903**	**0.838**	**0.869**
	TaggerOne	0.838	0.829	0.833

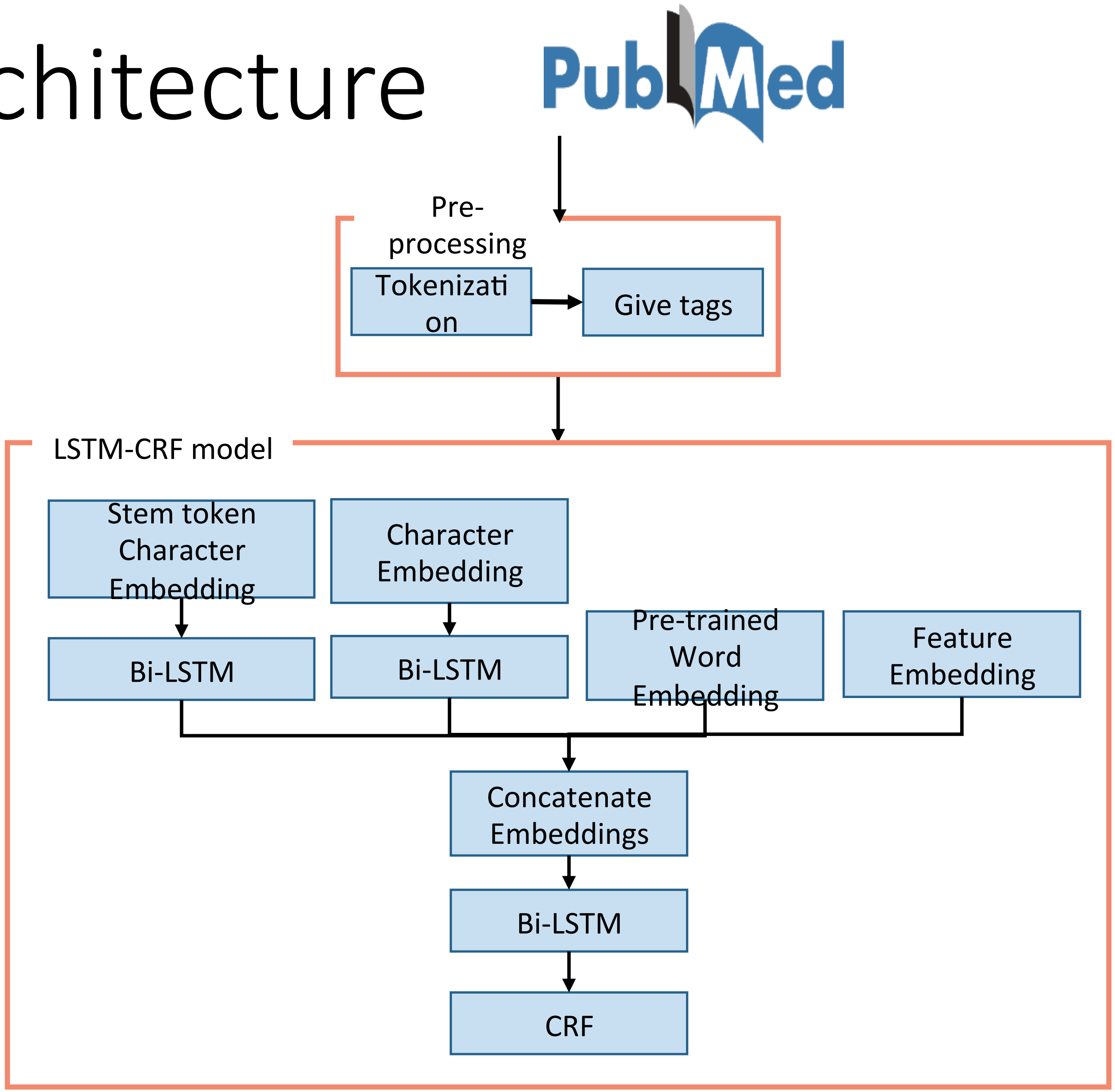
- Evaluation on chemical:

	Tools	Precision	Recall	F-score
NER	CDRnN	0.938	**0.892**	**0.914**
	TaggerOne	0.938	0.888	0.912
NEN	CDRnN	**0.8995**	0.879	0.889
	TaggerOne	0.879	**0.905**	**0.892**

Deep learning enhanced approach

- Stages of curating disease names:
  - Name Entity Recognition (NER)
  - Name Entity Normalization (NEN)
- To get less error propagation from NER to NEN, we use deepCRF to improve score of NER.
- Auto generation of CRF features
- Reference works:
  - Neural Architectures for Named Entity Recognition. Lample, Guillaume, et al., NAACL 2016.
  - End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. Xuezhe Ma and Eduard Hovy, ACL 2016

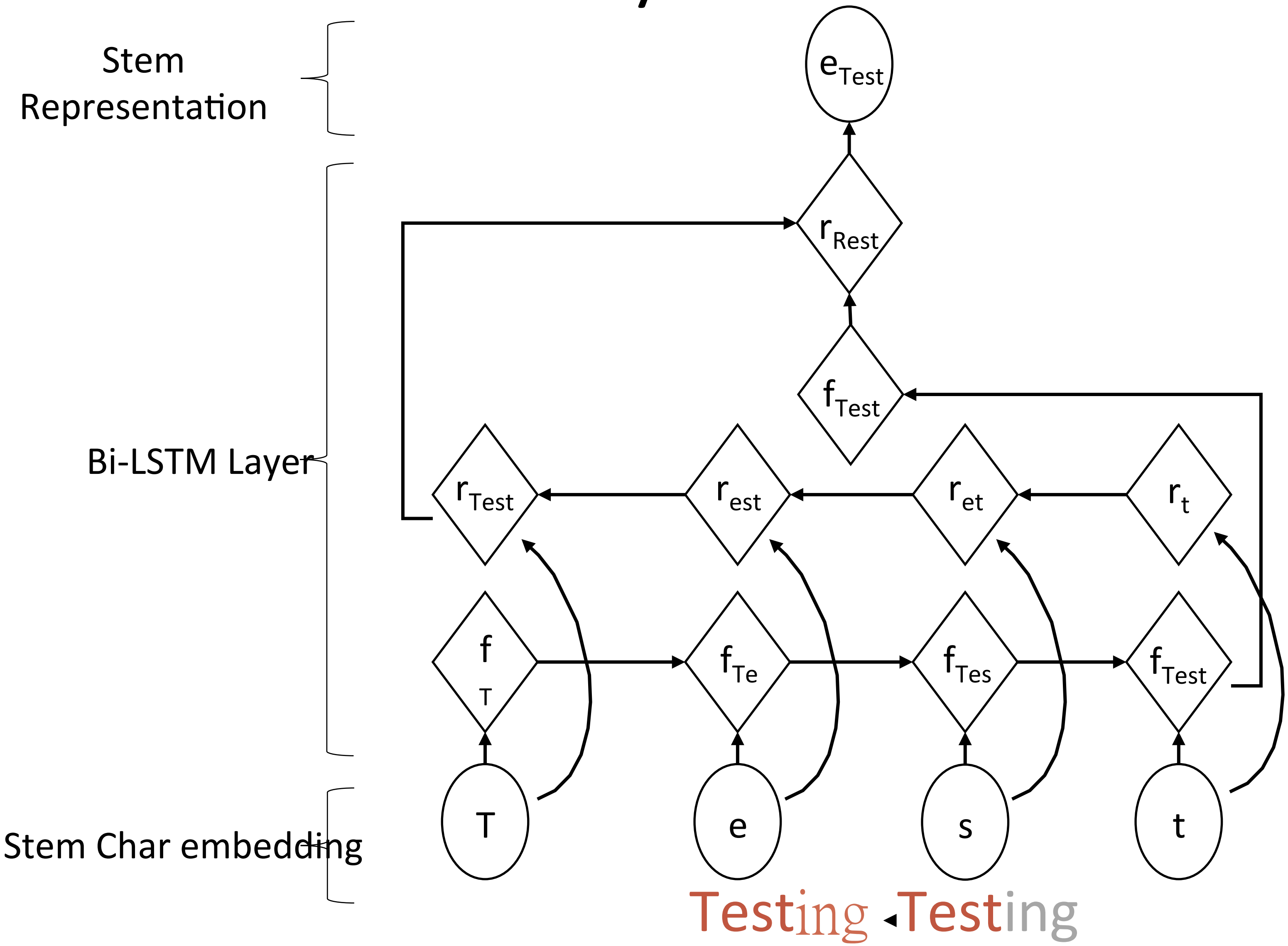
Architecture



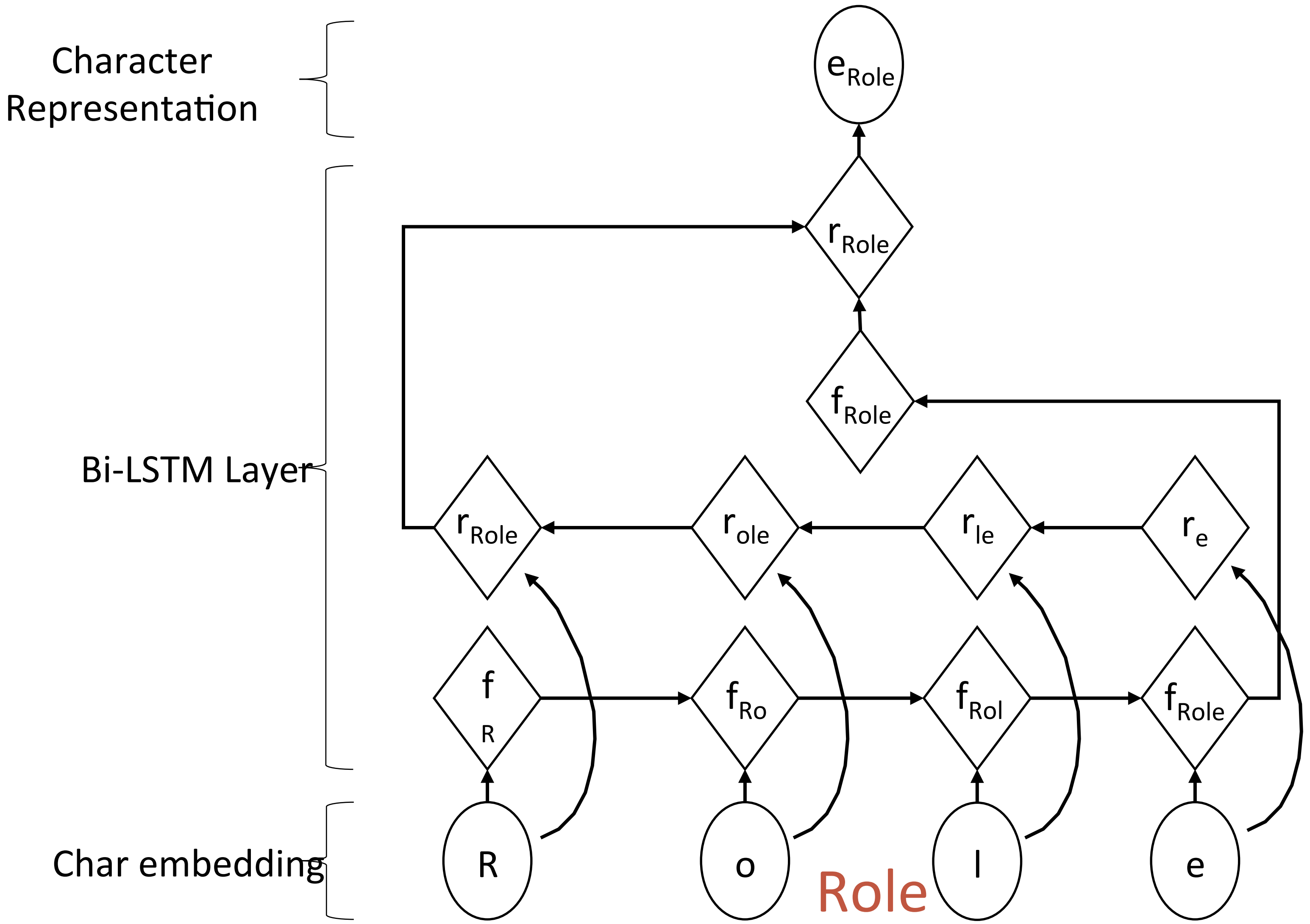
Character Layer/StemChar Layer

- Enhance the meaning of the word.
  - Character Layer:
    - The effect is similar to the prefix and suffix, but better.
    - We don't have to decide the length of the character.
  - StemChar Layer:
    - Obtain the stem of the word from Snowball Stemming Library.
  - Initialize the embedding with random vector.

StemChar Layer



Character Layer



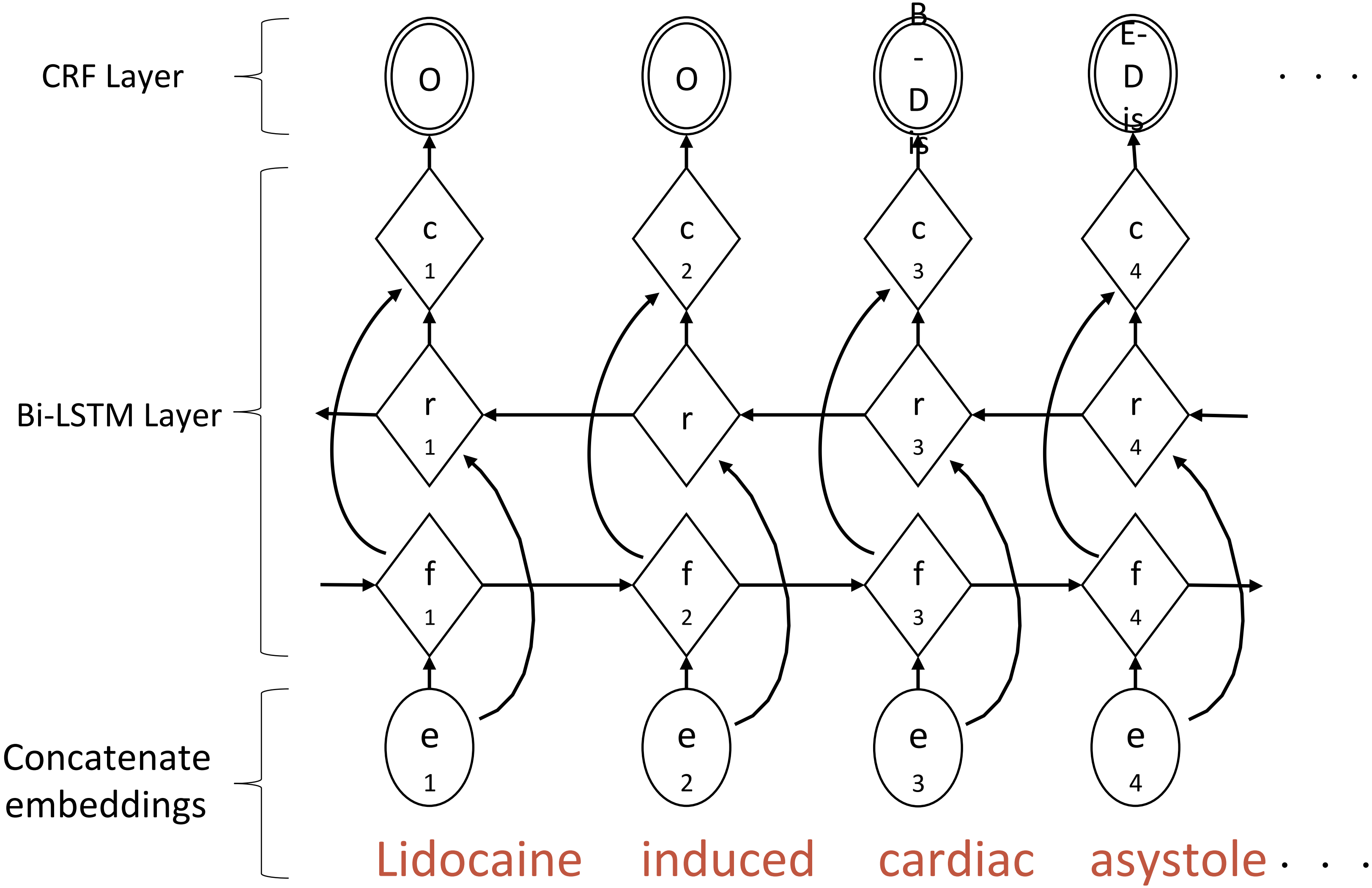
Feature Embeddings

- Word presentation:
  - Pre-trained Word Embedding : GloVe

- Feature Embeddings :

Disease ending word	• disease, tumors, tumor, cancer, damage, illness, illnesses, abnormality, abnormalities, tumour, abortion, abortions
Dictionary lookup	• CTD disease vocabulary (MEDIC) , NCBI disease corpus
Abbreviation	• BIOADI library

LSTM-CRF Layer

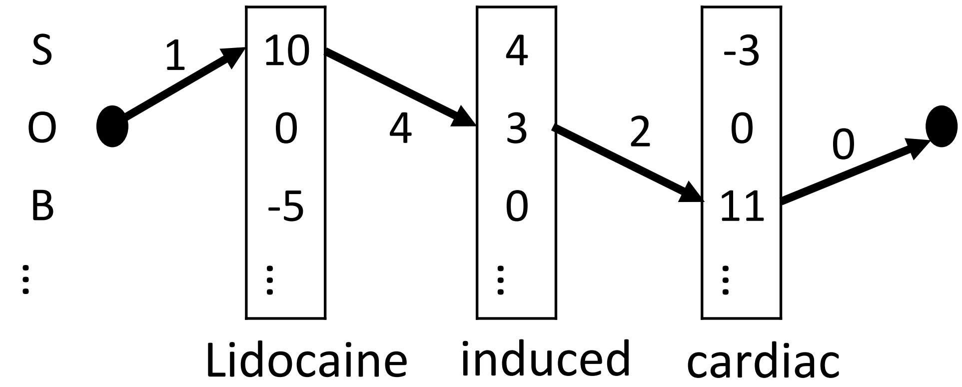


LSTM-CRF Layer

- LSTM layer -> CRF Layer:

Y: sequence of tags
X: a sentence
s(Y, X): score from the feature fu

  - We obtain every tensor of hidden_state from time_step in LSTM.
  - Input the tensor into the formula above to train the CRF model.



Data set description

- CDR corpus:

Consists of 1500 articles from PubMed, divided into three subsets with 500 for each.

Task Dataset	Articles	Chemical		Disease		CID
		Mention	ID	Mention	ID	relation
Training	500	5,203	1,467	4,182	1,965	1,038
Development	500	5,347	1,507	4,244	1,865	1,012
Test	500	5,385	1,435	4,424	1,988	1,066

Data set description

- CDR corpus:
  - PMID|t|Title
  - PMID|a|Text
  - PMID start_offset end_offset token_name type MeSH

```
354896|t|Lidocaine-induced cardiac asystole.
354896|a|Intravenous administration of a single 50-mg bolus of lid
of the sinoatrial and atrioventricular nodal pacemakers. The patie
to the development of bradyarrhythmias; and, thus, this probably r
354896  0    9    Lidocaine    Chemical    D008012
354896  18   34   cardiac asystole    Disease D006323
354896  90   99   lidocaine    Chemical    D008012
354896  142  152  depression   Disease D003866
354896  331  347  bradyarrhythmias    Disease D001919
354896  409  418  lidocaine    Chemical    D008012
```

Compare with other systems

System	Mention			Concept		
	Precision	Recall	F-score	Precision	Recall	F-score
Ours	93.23%	90.32%	**91.75%**	94.12%	83.75%	**88.63%**
AuDis[1]	87.55%	84.2%	85.9%	90.49%	84.26%	87.26%
Y.L. et al.[7]	89.61%	83.09%	86.23%	89.56%	85.71%	87.61%
Li et al. [5]	81.69%	84.75%	83.19%	-	-	-
M.H. et al. [6]	84.19%	82.79%	83.49%	-	-	-
Baseline						
TaggerOne [2]	85.2%	80.2%	82.6%	84.6%	82.7%	83.7%
LSTM-CRF [3]	78.96%	80.76%	79.85%	91.25%	74.54%	82.05%

Performance of each part

Run	Condition	NER F-score	NEN F-score
0	LSTM-CRF	82.82%	84.30%
1	Add Pre-train word embeddings	83.36%	84.70%
2	Add Feature embeddings	91.61%	88.68%
3	Add StemChar layer	91.75%	88.63%

Feature evaluation

Stem	• Snowball stemming
POS	• Stanford POS tagger
DTC	• disease, tumors, tumor, cancer, damage, illness, illnesses, abnormality, abnormalities, tumour, abortion, abortions
Vowel	• vowels change to “-” • “tumor” and “tumour” are turned into “t-m-r“
Prefix/Suffix	• Length from 1 to 5
DIC	• Dictionary lookup • CTD disease vocabulary (MEDIC) , NCBI disease corpus
Abbreviation	• BIOADI
Terminology	• disease terminologies • body part • human ability

Feature evaluation

Removed feature		Precision	Recall	F-score
All	Mention	91.91%	89.89%	90.89%
	Concept	93.76%	83.14%	88.13%
-Prefix/Suffix	Mention	91.32%	88.99%	90.14% (↓ 0.75%)
	Concept	93.51%	83.45%	88.19%
-Vowel	Mention	91.75%	90.30%	91.02% (↑ 0.13%)
	Concept	93.22%	83.70%	88.20%
-Stem	Mention	93.01%	90.34%	91.66% (↑ 0.77%)
	Concept	93.76%	83.90%	88.55%
-POS	Mention	92.14%	89.94%	91.03% (↑ 0.14%)
	Concept	93.97%	83.95%	88.68%
-Terminology	Mention	91.99%	89.64%	90.80% (↓ 0.09%)
	Concept	94.14%	83.24%	88.36%
-DTC	Mention	93.18%	87.77%	90.39% (↓ 0.5%)
	Concept	94.30%	83.24%	88.43%
-Dictionary lookup	Mention	83.68%	83.70%	83.69% (↓ 7.2%)
	Concept	90.96%	77.96%	83.96%
-Abbreviation	Mention	92.48%	85.66%	88.94% (↓ 1.95%)
	Concept	93.76%	84.00%	88.61%

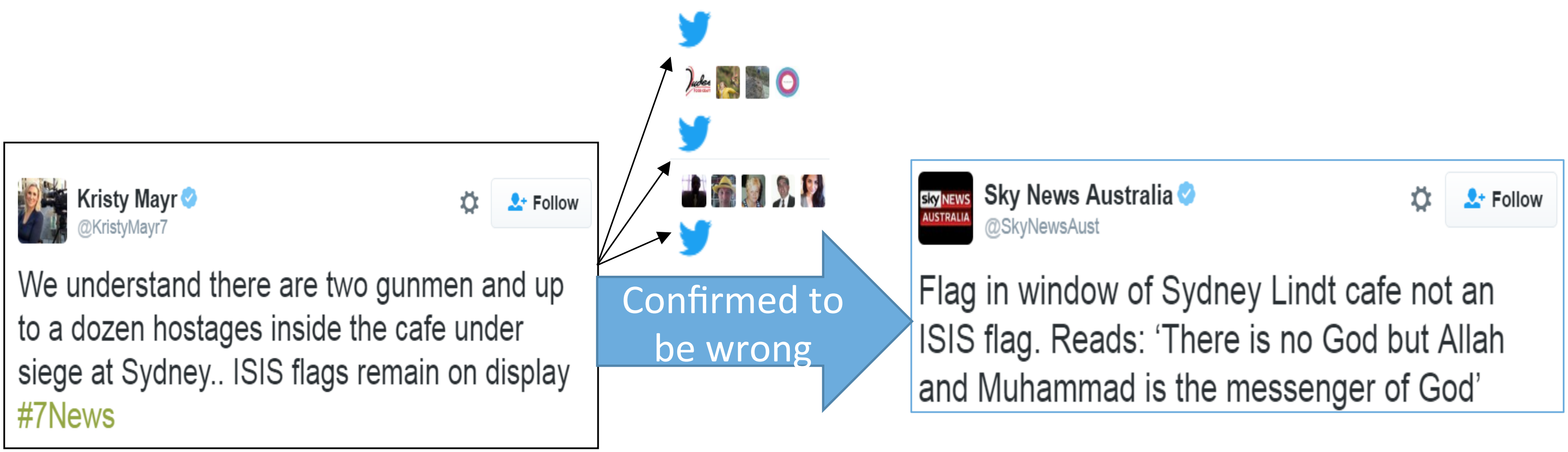
Performance of other dataset

NCBI Disease corpus	System		Precision	Recall	F-score
Trained by Trainset	Ours	**Mention**	96.13%	95.83%	**95.98%**
		Concept	93.39%	95.58%	**94.47%**
	TaggerOne [2]	**Mention**	85.1%	80.8%	82.9%
		Concept	82.2%	79.2%	80.7%
	Y.L. et al.[7]	**Mention**	90.72%	74.89%	82.05%
		Concept	88.73%	77.30%	82.62%
Trained by random 70% pages from all the corpus	Ours (Avg. 3 times)	**Mention**	97.39%	96.88%	**97.13%**
	M.H. et al. [6]	**Mention**	85.31%	83.58%	84.44%

Rumor detection / Stance classification

Introduction – *Background*

- What is rumor?
  - Rumor is a **controversial, fact-checkable** statement.
  - We focus on non-factual (false) claim.
- Unconfirmed rumors usually sparks discussion before being verified



Introduction – *Background*
Spread of Rumor

- Retweet of a false rumor in Event Sydney Siege spread all of the world.

Source from Sydney

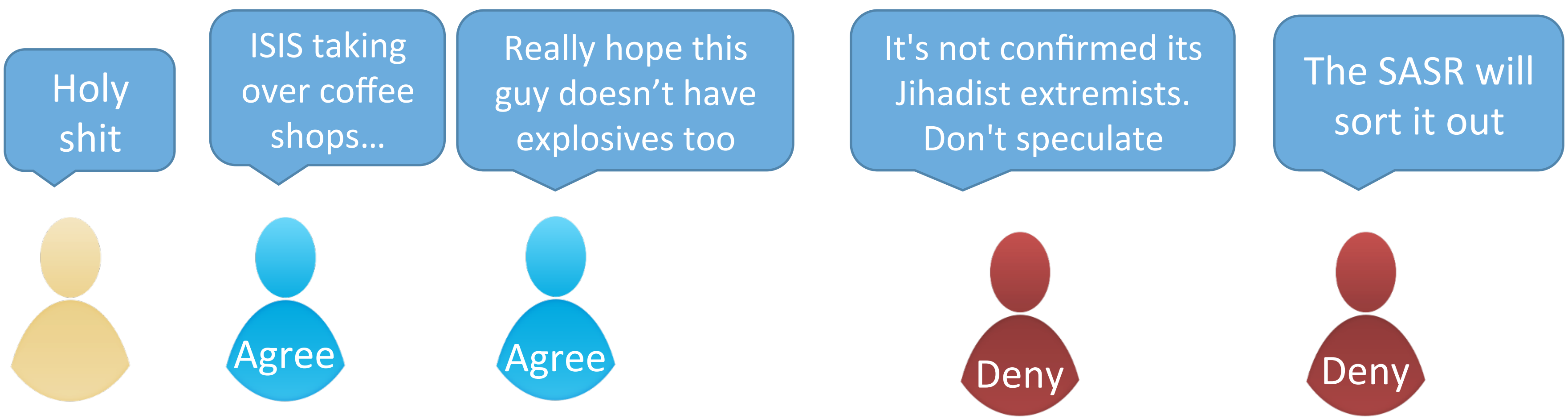
From where people share the Tweet.
Font size correspond to quantity.

Introduction – *Background*

Rumor trustworthy

- How likely will you believe in the message?

And Now?



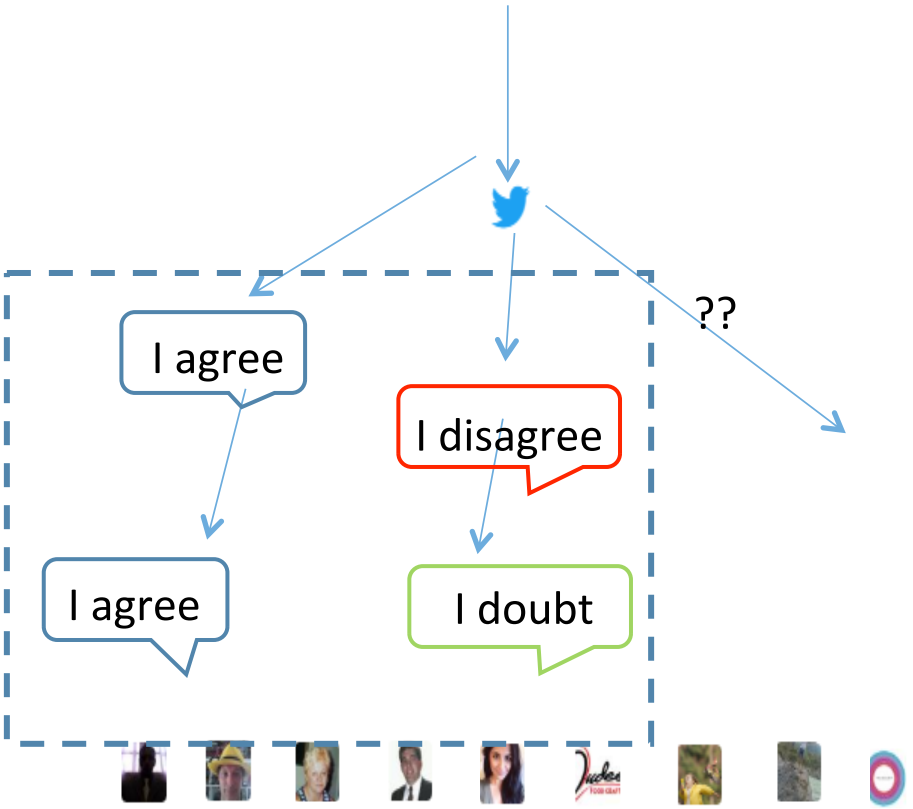
Trust of messages to people will change while more replies revealed.

Introduction – *Motivation*

- **Twitter** is a valuable source for journalists
  - First hand
  - Real-time (External data is unavailable)
- **Tweets vs. Traditional news press**
  - Text
    - Short text(140 characters)
  - Feature
    - Users interactive
    - Messages propagation

Introduction – *Motivation*

A Twitter response structure



Can we predict the stance of next user by the previous messages?

Introduction – *Motivation*

- It is likely to have **conflict opinions** on rumors.
- **Users may change their perception** of credibility toward a post depending on the new evidence uncovered.

u1: AFP reports there are 2 dead and 5 hostages being held in the Kosher store in Eastern Paris; separate incident to charlie hebdo shooters. **[Support]**

u2: Not so much separated. There're links between them. **[Deny]**

u1: meant it in the sense that it's different person/people. Links between]them unclear at the moment, lots of rumours.

u3: Yes lots of mixed info coming out. Anyone else here about potential 08 prison connection btwn planners? **[Query]**

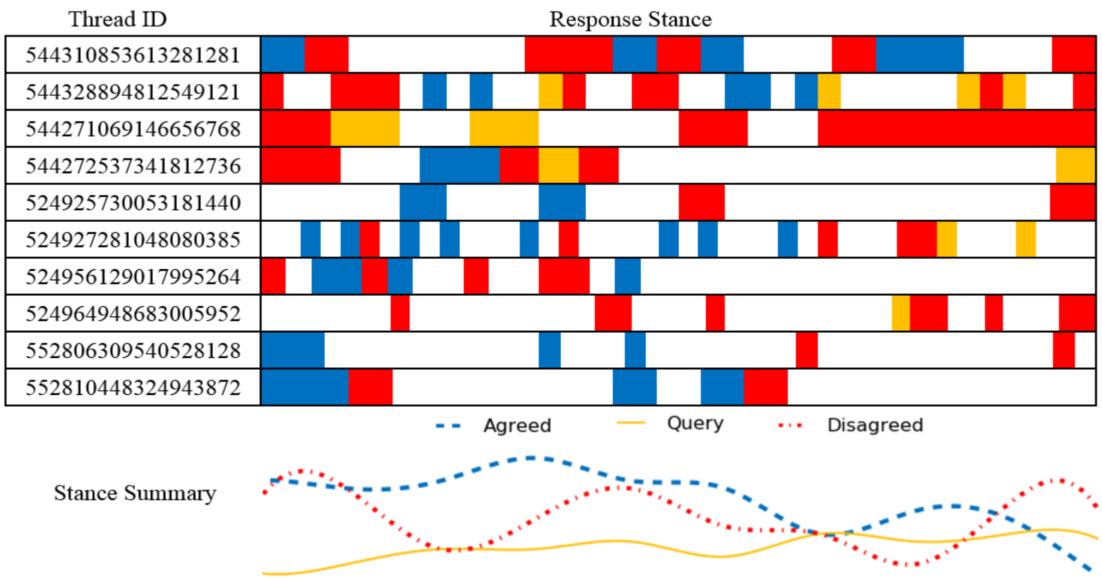
u4: Speaking of https://t.co/lCQ0qFyePh **[Support]**

u2: Honestly, we need to wait official statement and investigations. By now, we all talk about rumors. :) **[Comment]**

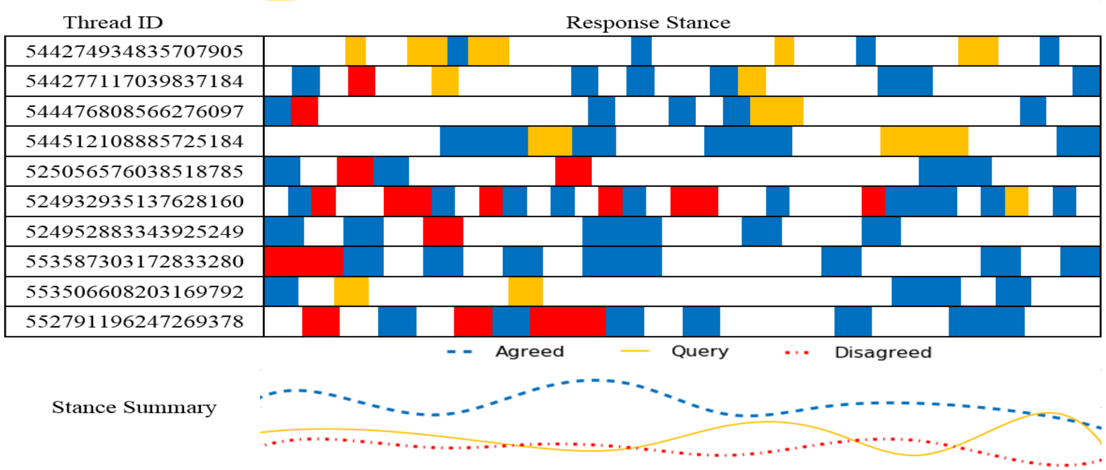
u5: Yes there is a link. It was confirmed officially today. **[Support]**

Introduction – *Motivation*

The quantity of stance at the different time interval



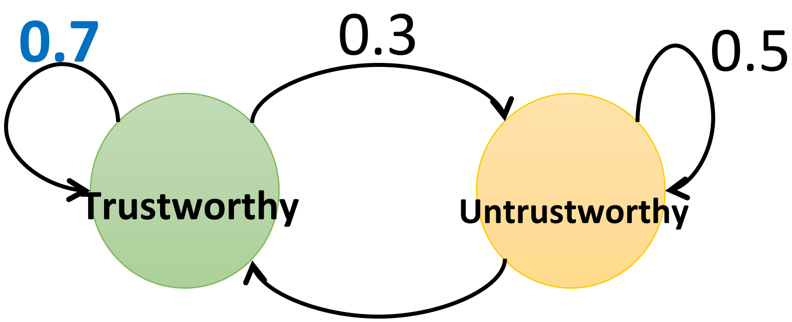
False rumor stance in thread



Normal message stance in thread

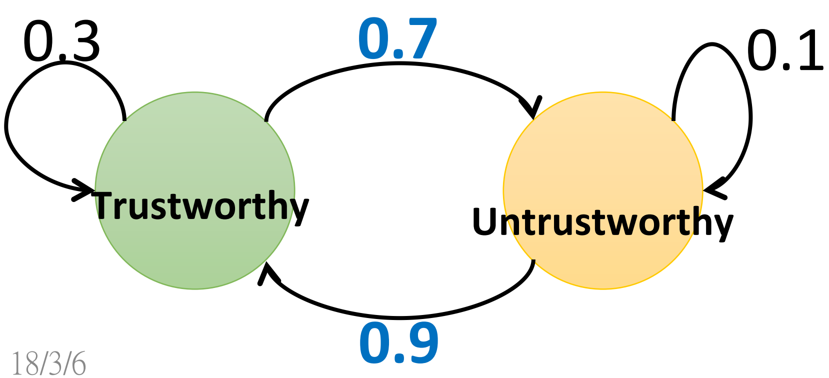
Introduction – Motivation

- Most work focus on finding new features.
- From the probabilistic view, we enhance the performance and tracking states of messages which are invisible before.



Transition probability of a normal message

Trustworthy state is stable



Transition probability of a false rumor

Untrustworthy state is not sustainable

Problem definition

Given Twitter data about an event, including message with meta data, replies and retweets.

	Stance Classifier	Rumor Verification Classifier	Credibility change by Propagation
Goal	Categorize the replies into one of the SDQC categories by reply-source pairs.	Label the rumor $\in$ {true, false} Normalizing scores from 0 to 1.	Improve messages credibility as well as their sub-event credibility.

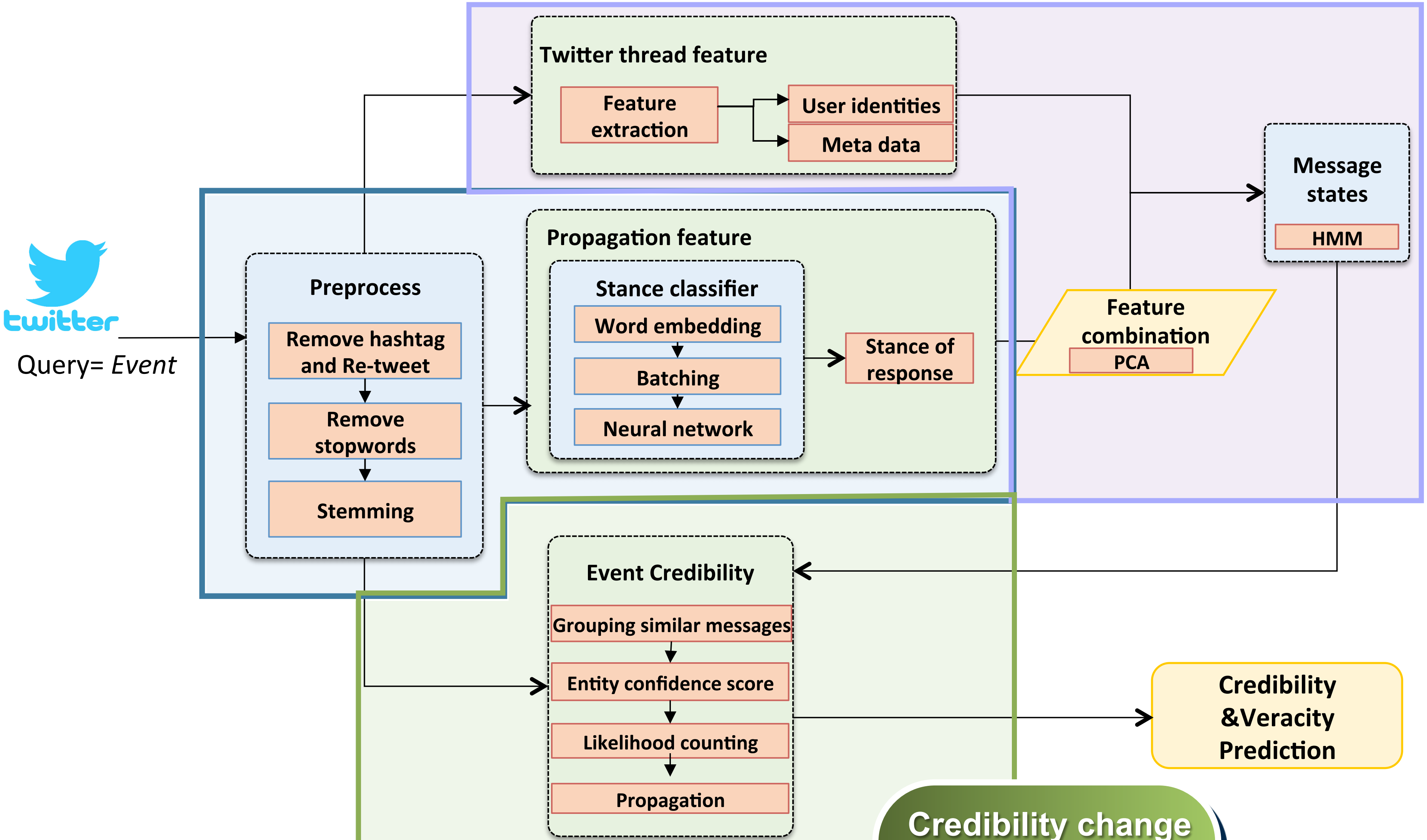
Problem :

- Data sparse/screw
- Difficulty of understanding the semantics

Stance classifier

Rumor verification classifier

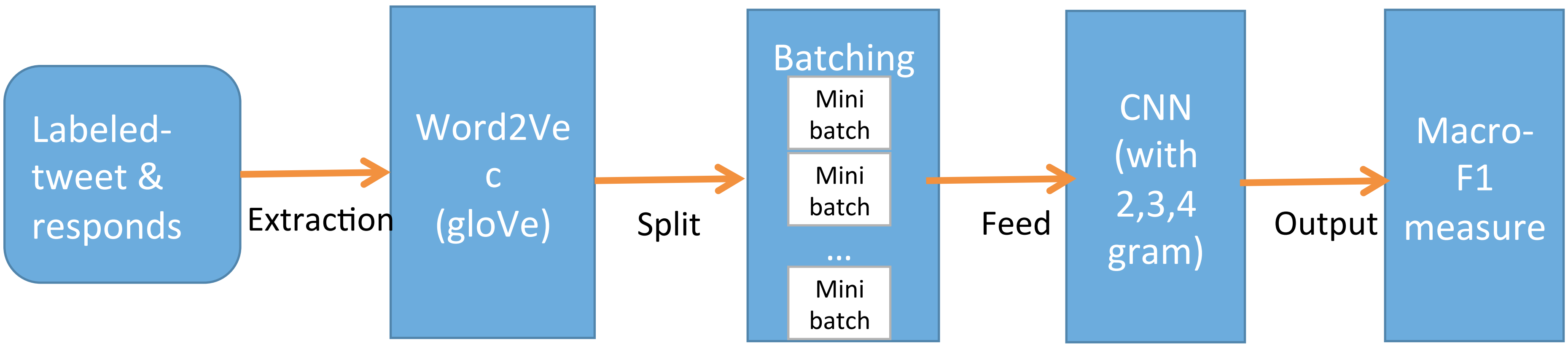
Method--*Overview*



Credibility change by propagation

Method – *Stance classifier*

Preprocessing → embedding → batch → training → testing → evaluation



Method – *Stance classifier*

Deep Learning Model

Why deep learning

Training examples can contain errors:

- Neural network models are robust to errors

Automatic feature extraction

- Backpropagation

Convolutional Model

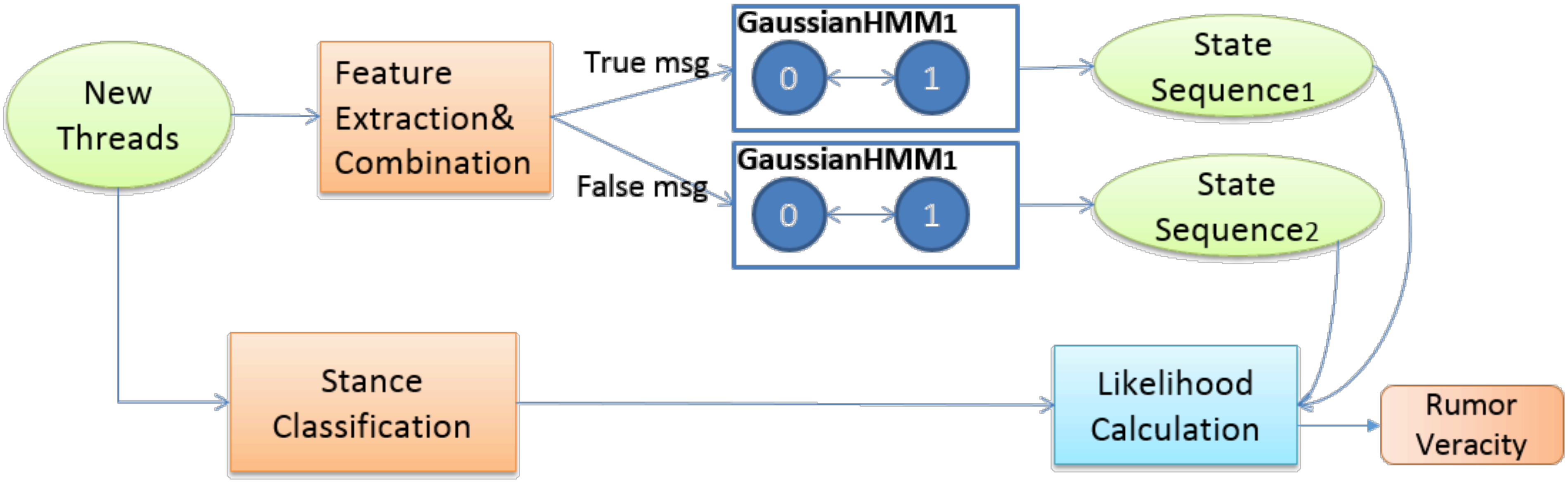
Sliding windows over 2-5 words at a time

Difficulty

Large amount of parameters to select

Method – Rumor Verification Classifier

- We trained two HMMs based on different observed data
  - False rumors
  - True rumors



Method – Rumor Verification Classifier

- Feature extraction

User Feature

- Influence:
  - #Followers (int)
  - #Friends (int)
  - #Tweets (int)

Message Feature

- Time (int)
- User mention(binary)
- Links(binary)

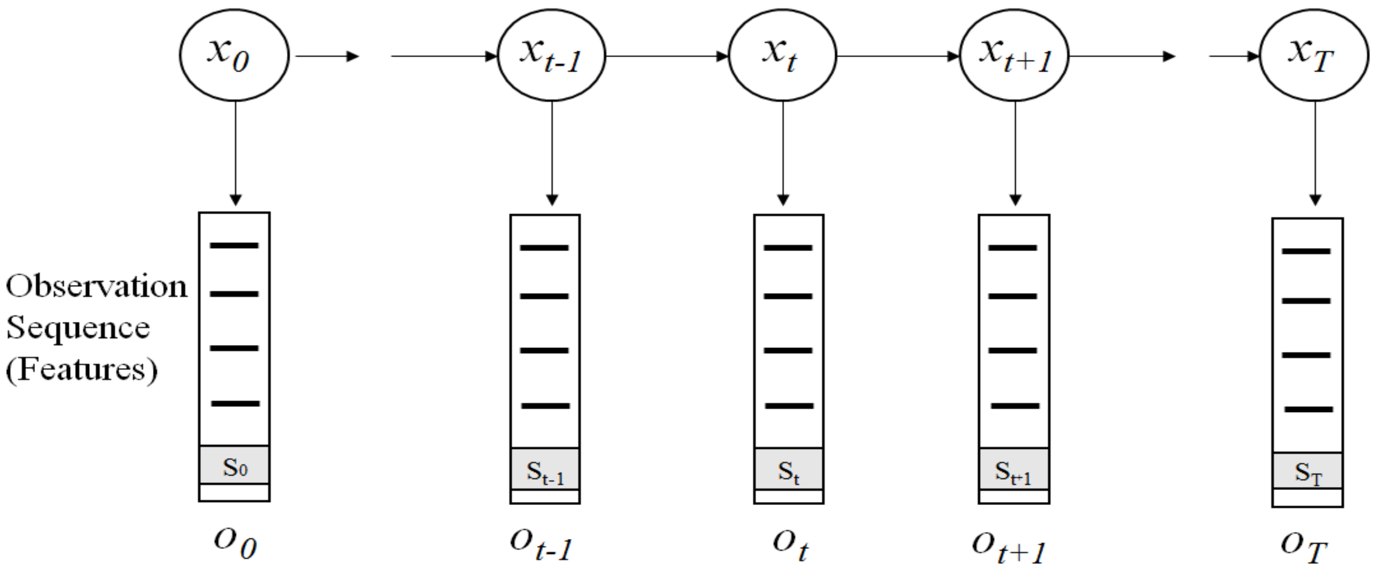
- #Comments(int)
- #Retweets(int)
- #Like(int)

- Emotions: POS(int)/NEG(int)
- Punctuations (binary)
- Stance(SDQC)
- Retweet(binary)

Feature Combination (PCA)

Method – Rumor Verification Classifier

- ***Markov Trust Prediction Models***



$$P(F_{T+1}|\lambda) = \pi B_{x_0,f_0} A_{x_0,x_1} B_{x_1,f_1} \cdots A_{x_{T-1},x_T}$$

A : state transition probabilities

$B^{x,y}$: observation probability is derived on **the past outcomes of** and the associated **feature sets**.

Experiments – *Datasets*

[1] Zubiaga A, Liakata M, Procter R, Wong Sak Hoi G, Tolmie P (2016) Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. PLoS ONE 11(3): e0150989. doi:10.1371/journal.pone.0150989.
[2] Jin, Zhiwei, et al. "News credibility evaluation on microblog with a hierarchical propagation model." Data Mining (ICDM), 2014 IEEE International Conference on. IEEE, 2014..

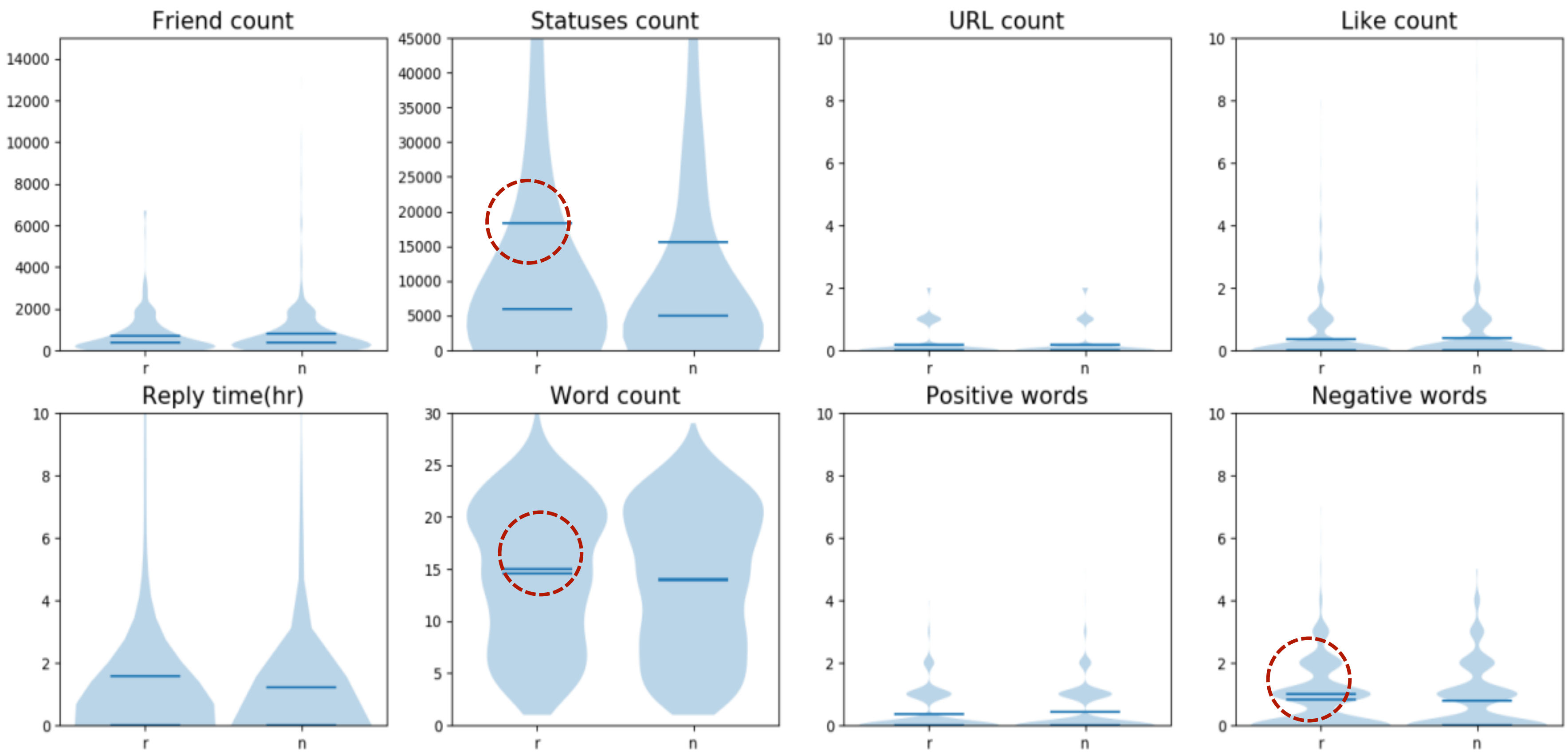
- Twitter Dataset [1]

Event	Charliehebdo	Germanwings-crash	Ottawashooting	Sydneysiege
Misinformation	12	12	10	13
Total tread	74	25	58	71
Avg length of response	16.20	11.69	13	16.51
Avg length of Retweet	121.2	87.1	145.3	166.8

- Weibo Dataset [2]—Flight MH370 Lost Contact

False Rumor	#Annotated post	#Annotated replies	#Total post
飛機已找到 (Lost planes found)	5	351	242
降落在Nanming機場 (Landing on Nanming airport)	21	1128	328
假護照恐怖攻擊 (Fake passport and terrorist attack)	8	302	52
乘客是恐怖份子(A passenger is terrorist)	8	213	105
Normal message	#Annotated post	#Annotated replies	#Total post
波音在協助搜索 (Boeing Corp. is assisting searching)	3	286	91
飛行路線曝光 (The route of flight revealed)	6	217	269
在西南方向 (On the South-East side)	2	44	29
慰問家屬(Condolences to family members)	8	351	497

Experiments –
Value distribution of features between rumor microblogs and normal ones



Negative words are more common on the false rumor message

Experiments – *Result of stance classification*

- We compare our deep learning model with other classifiers From Lukasik' s paper[1]

	Ottawa		Ferguson	
	Acc	F_1	Acc	F_1
Majority vote	61.51	19.04	66.86	20.04
SVM	64.58	35.39	66.86	20.04
GP	62.28	42.41	64.31	32.9
Lang. model	53.2	42.66	49.56	34.35
NB	61.76	40.64	62.05	31.29
CRF	64.58	33.07	67.35	28.11
HP Approx.	67.77	32.29	68.44	25.99
HP Grad.	63.43	42.4	63.23	33.14
CNN	61.74	**44.9**	62.31	36.49
CNN(word2v)	59.61	38.87	63.03	**39.48**
RNN(word2v)	52.49	38.66	51.49	32.52

In Macro-F1 , CNN (dynamic embedding) and CNN(word2v) perform better than others.

[1]M. Lukasik, P. K. Srijith, D. Vu, K. Bontcheva, A. Zubiaga, T. Cohn. **Hawkes Processes for Continuous Time Sequence Classification: an Application to Rumour Stance Classification in Twitter**. ACL. 2016.

Experiments – *Stance Classification*

Statistical details of each class

Stance	Precision	Recall	F_1 score
Support	0.19	0.20	0.19
Deny	**0.31**	0.07	0.11
Query	**0.58**	0.45	0.51
Comment	0.78	0.85	0.81

Comment : Take the most part

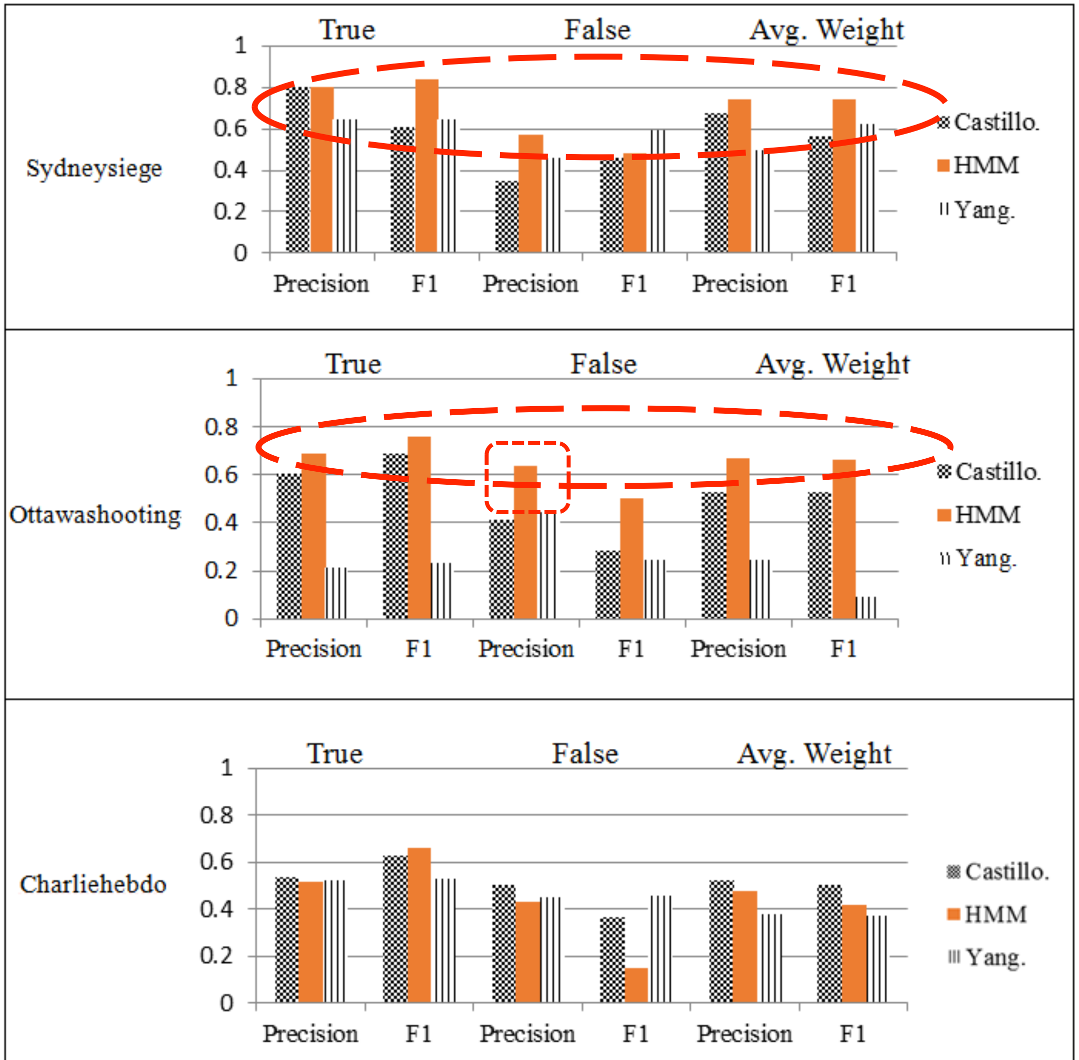
Query: Easy to detect because the features are more obvious

Deny : Negative words is a good feature

Support: Hard to detect.

Experiments –*Rumor Verification*

Performance compared with baseline by F_1 and precision



Castillo:

- J48 decision tree
- 15 best reported features

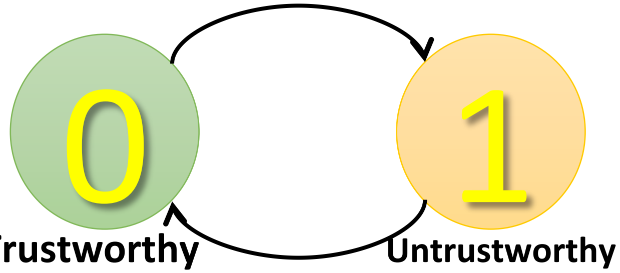
Yang:

- SVM classifier
- 19 reported features

F. **Yang**, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," SIGKDD 2012
c. **Castillo**, M. Mendoza, and B. Poblete, "Information credibility on twitter," WWW 2011

Experiments –*Rumor Verification*

Case study: Difference Pattern between normal HMM vs. rumor HMM



Trustworthy state is more often to be observed in the normal HMM than rumor HMM

Rumor HMM	Normal HMM	Stance	Messages
0	1	Agree	"@7NewsSydney: The Sydney cafe siege may be part of a larger plot. @Y7News #MartinPlaceSiege http://t.co/d07MMUcpKT http://t.co/PL06PUyfDF"
…			
1	0	Comment	@RPW260567 @jewelsparkle3 @7NewsSydney @Y7News But Doyle coming up! Ugh!
1	0	Comment	@ZZiillaa @jewelsparkle3 @7NewsSydney @Y7News She'll repeat everything as if we're all stupid.
1	0	Disagree	@7NewsSydney @Y7News Unless you've got evidence of this...shut it. Sick to death of my emotions being manipulated for money. #parasites
1	0	Disagree	@7NewsSydney @Y7News Stick to the facts and stop trying to scare thr australian puplic!
1	0	Agree	.@7NewsSydney @HaynesParker1 @Y7News Duh, of course it's "part of a larger plot", #Islamic #terrorists hate ppl of all religions
1	1	Disagree	Yes: https://t.co/tt33b9GgH2 Note: unconfirmed @7NewsSydney @2kdei @Y7News @PzFeed @BBCWorld @ABC
1	0	Agree	RT "@7NewsSydney: The Sydney cafe siege may be part of a larger plot. #MartinPlaceSiege http://t.co/wZ2bmrC2AA http://t.co/qNwgDycEWn"
0	0	Comment	@7NewsSydney @2kdei @Y7News Yes, a red herring, or a feint...
1	1	Comment	@7NewsSydney @Y7News please dont scare people by speculating whats gonna happen.. My prayers to all the people taken hostaged.. :(
1	0	Disagree	@7NewsSydney @Y7News your coverage is shameful , stick to the facts ...
1	0	Agree	@7NewsSydney @norsepast @Y7News http://t.co/YQTxEeSQem
0	1	Agree	"@7NewsSydney: The Sydney cafe siege may be part of a larger plot. @Y7News #MartinPlaceSiege http://t.co/D6ucJKSuNv http://t.co/OSFD2ZhPbz"
0	0	Comment	@7NewsSydney @Fitzzer777 @Y7News Yes, it's called "Islam."
1	0	Comment	@7NewsSydney @Y7News in sure Obama will tell us just #workplaceviolence
1	1	Disagree	@Melstar71 @7NewsSydney @Y7News their entire coverage has been a joke. See the uni professor explain hypotheticals in hostage situations?

Experiments –Hierarchical Propagation

Weibo Case study

Event credibility (iterative)	Event credibility (voting)	Message credibility	Post
0.04	-0.04	-0.5	#马航飞机失联#【成都晚报记者称马航客机是【降落！！！！！】】原定于13：30召开的发布会，在超过原定时间50多分钟后终于开始。马航总经理现场讲话，航班已证实在越南南宁[音]地区降落。via成都晚报'
		-0.24	成都晚报刚刚在官方微博上称“马航总经理现场讲话，航班已证实在越南南宁[音]地区降落。” 而不少媒体连线前方记者，皆称发布会上没有关于飞机已经降落在越南机场的消息。[伤心]'
		0	【马来西亚官员：尚未获得越南方面消息】（宋宇 葛菁）马来西亚国防部长兼代理交通部长拿督斯里希沙穆丁表示，希望不会有任何猜测。飞机降落在南宁的消息并不属实；与越南海域相关的消息，也不属实。他表示将每两小时公布一次消息，将只公布可靠的官方消息，以免引发推测。马来西亚官员：尚未获得越南方面消息'
		0.5	#马来西亚航班失联#【辟谣】 有消息称，马来西亚失去联系的MH370航班因雷达故障，降落在一个叫做Nanming的机场。记者目前已经与南宁机场管理集团取得联系，没有该航班降落的消息，因此上述机场很可能在国外。依据此前民航局空管局的消息，该飞机没有进入中国管制区。祈福，平安！据中国经营报'

Sentiment-aware Chatbot

-- In view of training corpus collecting

Sentiment aware chatting

- 今天是要吃什麼？
  - 根本沒差好嘛

- 想去吃什麼?
  - 跟你出去吃什麼都可以

Sequence Generation-Sample output

Some sample output from machine generator

```
PRED SCORE: -7.0794
SENT 46024: 雇 用 殺 手 殺 自 己
PRED 46024: 不 用 叫 人 家 歸 於 歸
TGT 46024: 不 如 去 找 人 同 歸 於 盡
PRED SCORE: -6.1666
SENT 46025: 現 在 還 有 人 用 黑 莓 機 嗎
PRED 46025: passport 花
TGT 46025: passport 潮 -
PRED SCORE: -4.8464
SENT 46026: 有 沒 有 小 火 鍋 的 八 卦 `
PRED 46026: 幹 這 樣 也 一 篇
TGT 46026: 自 己 買 料 回 來 自 己 煮 啊
PRED SCORE: -7.8754
SENT 46027: 妹 妹 森 77 了 怎 摸 半 -
PRED 46027: 臭 酸 味 滾 喇 幹
TGT 46027: 酸 臭 癢 廢 魯
PRED SCORE: -3.4965
SENT 46028: 最 近 一 次 哭 是 為 什 麼 `
PRED 46028: 初 音 是 民 國 人
TGT 46028: 初 音 是 個 軟 體
```

Training data Collecting / generation

- Template-based generation
- Weibo
  - Emote symbol
- PTT
  - Board characteristics
  - User annotation

Sentiment Analysis - Template Generator

Generate some **training** data by template

- Example:

  [S1][Time]被[S2][SV] (了)

  我 昨天 被 同學 欺負 了

  我 上禮拜 被 老師 罵

  我 被 雷 了

  ...

Sentiment Analysis - Corpus

Explore WEIBO chat corpus

- I assume that the **emotes** can imply the utterance's sentiment
  - For example:
    - 红色炸弹一下子全来了,我没穿防弹衣,[泪]
    - 非常好![呵呵]
    - [汗]打了加起来差不多一百分钟[嘻嘻]有惊喜没?
    - 今天心情特别好呀特别好~~[嘻嘻]
    - 我疯不来...这么早你居然给我堵车...[怒][抓狂]

- There are about 530,000 pairs have emotes
- We had selected 173 kind of emotes(#occurance > 180) as our labels
- Top 10 #occurance emotes
  - 泪 哈哈 偷笑 嘻嘻 抓狂 爱你 怒 汗 挖鼻屎 鼓掌

Sentiment Analysis - Corpus

- Number of Tags
  - Possible sentiment:
    - Happy: 哈哈/偷笑/嘻嘻/抓狂/爱你/鼓掌/doge/酷/good/笑cry/呵呵/微笑/太开心/给力/赞/耶/亲亲/笑哈哈/噢耶/得意地笑/开心/好棒/大笑/胜利/好喜欢
    - Sad: 泪/晕/衰/生病/可怜/悲伤/委屈/眼泪/失望/伤心/泪流满面/大哭/崩溃/感冒/悲催/难过/心碎
    - Angry: 怒/鄙视/怒骂/闭嘴/愤怒/斜眼/别烦我/白眼
    - Impatient: 汗/挖鼻屎/哼/打哈气/困/浮云/黑线/挖鼻/囧/懒得理你/右哼哼/左哼哼/打哈欠/傻眼/巨汗/抠鼻屎/
    - Surprise: 吃惊/惊恐/震惊/紧张
    - Shame: 害羞/可爱/羞嗒嗒/脸红/不好意思/羞羞

PTT emote source

Board	Main	Tag	Size
Broken_heart	Neg	Sad	23,300
happy	Pos	N	23,727
Hate	Neg	Angry	76,141
Lucky	Pos	N	18,000
prozac	Neg	Sad	36,201
Sad	Neg	Sad	26,017

Which board is really matters?

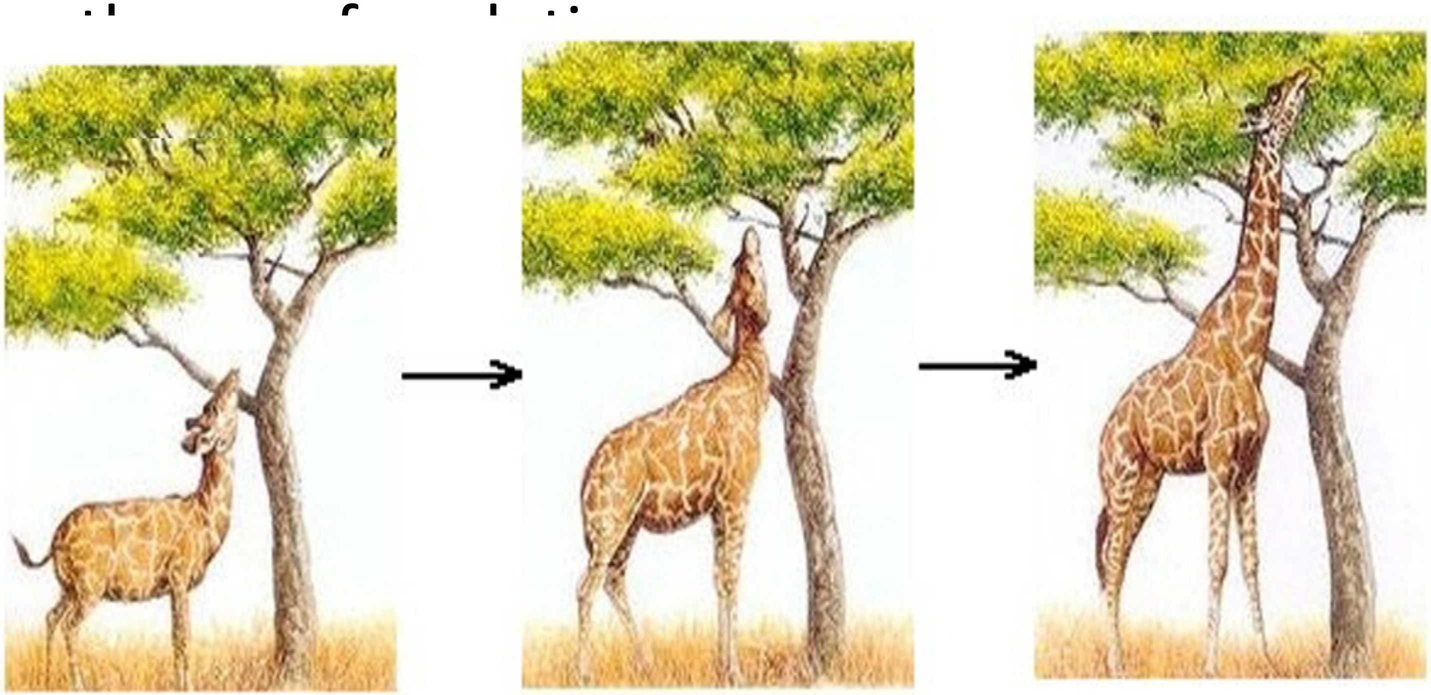
- Wish, Lonely, Marginalman, Love
  - 明天放假，希望不要下雨 (POS)
  - 兵單快點來啊 (NEG)
  - 希望大家都能幸福 (POS)
  - 我許了一個願　(NEG)

Stance / sentiment classification

Type	Dev set	Val set	Epoch	Note
CNN-Char	~96 %	84.432%	15	lr=1e-4, tri-gram, batch=768
CNN-Char-NoStopWords	~95 %	83.672%	15	lr=1e-4, tri-gram, batch=768
CNN-Char	~97 %	83.811%	15	lr=1e-3, tri-gram, batch=768
CNN-Char	~97.5%	83.807%	25	lr=1e-4, quar-gram, batch=512
CNN-Word-NoStopWords	~96%	82.569%	15	lr=1e-4, tri-gram, batch=768
CNN-Word	~97.5%	82.573%	25	
RNN-Char	95.703%	83.937%	25	lr=1e-3, batch=512
RNN-Char-NoStopWords	96.120%	83.033%	25	lr=1e-3, batch=512
RNN-Word	96.484%	83.273%	25	lr=1e-3, batch=512
RNN-Word-NoStopWords	95.312%	82.197%	25	lr=1e-3, batch=512

GANs

- The principle of GAN is to rely on the antagonism between the generator and discriminator to make each other stronger, which is a concept similar to t



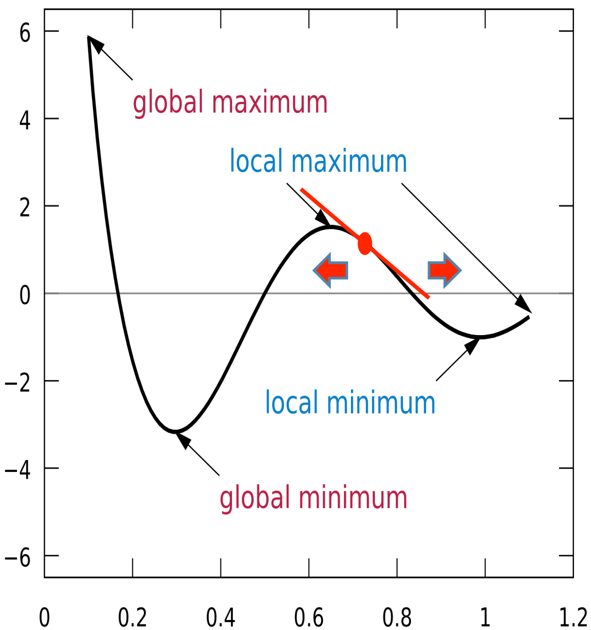
The first version

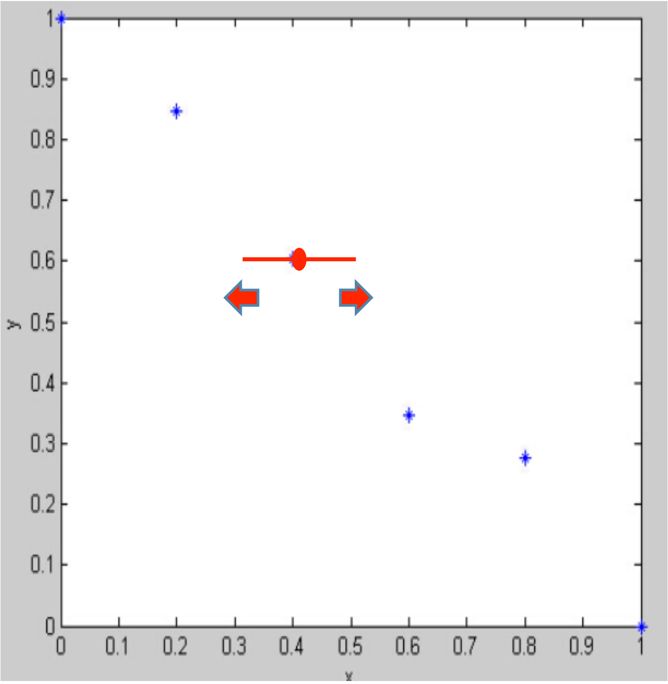
The second version

The final version

The plight of GANs in the field of NLP

- Discrete domain datasets which can not be differential.



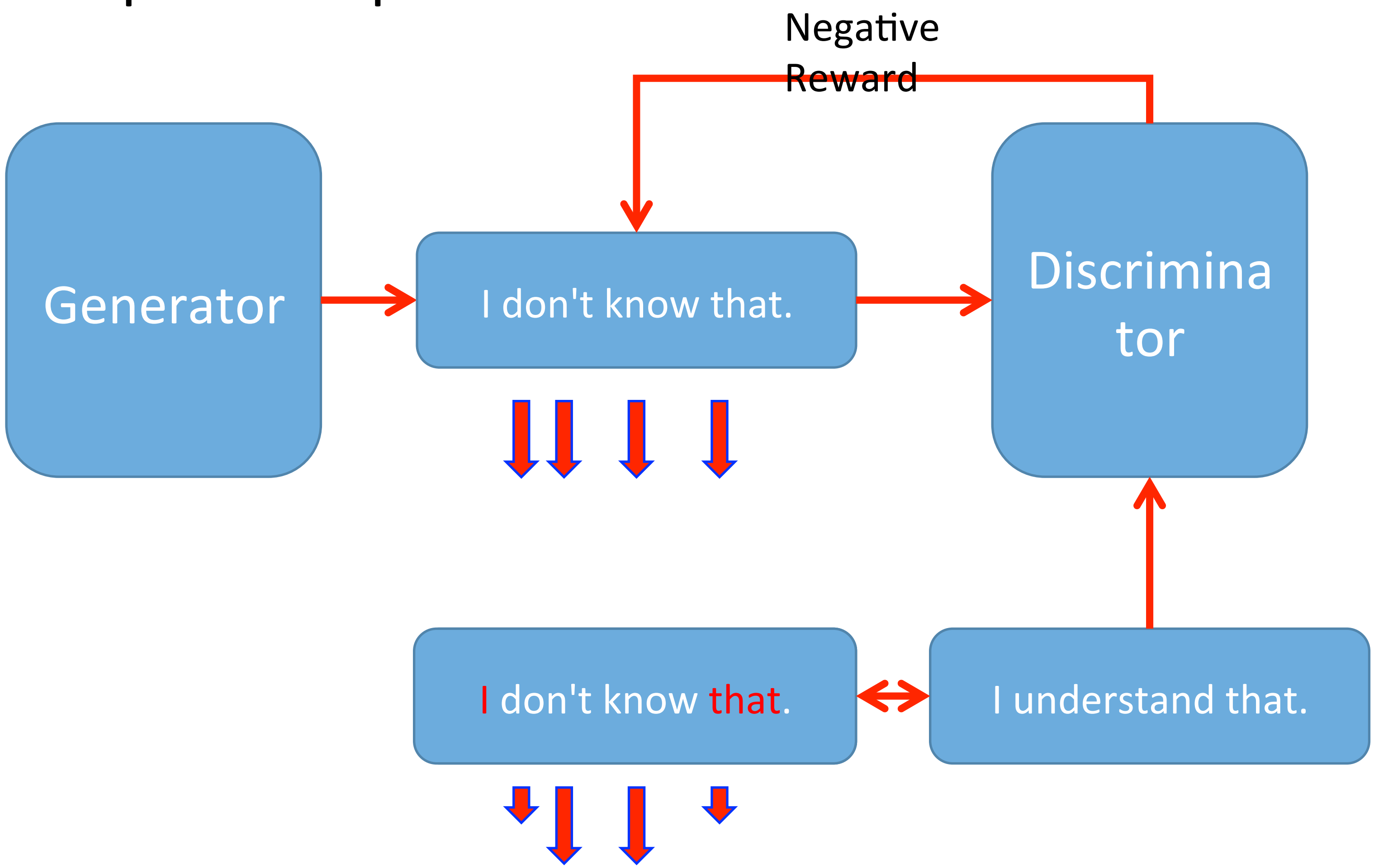


$$W = W_{old} + \alpha * gred$$

$$W = W_{old}$$

The plight of GANs in the field of NLP

- A simple example



Experiments

- Sentences generated by WGAN

```
,Aöäép5mBmATVmfm&Jtï
?A@è-´A@mm1éTfLä‘,ïm
päz’RJmcASp8Jr*Aiáá2
AmT8ää*B)f#fé&ä)m♥Mü
éég“J2:äé–c’T8hf+(üL
rcp@fbmY5z2AdgAf]A`á
@)-AcMmfMJiéJg*m‘„f.
Éicmé]äJ:’bzgJu$(Mfn
cpmñAl„(éff*CGTJfV8E
uPbTTlfTYä´f:ÀU2gAm"
@iEä‘ägä@‘ä8D2(äAä`A
B8chéäBudJTb=ú$úmTAó
‘dAó„äfd22f‘2ffrAèéI
éJämC‘pS´mA"éu`mmä'ö
IM1‘mlz&ät&óáAA]5:Ut
r-7Écbm@J:AazJ2T2Smé
é([éRmTÉé5c'Jä2m*lJ8
dffr2ä*"@J`ÀémäpégA-
‘cM‘C–]fcéJAr@fÖA8´%
```

epoch 0

```
ooo ohe t  ohe thoo
o too ooootoooooht t
ho totooho e tototoo
o tooht tothshe t t
 t oooooohe tho e eo
he ththt e tht the t
e tooho tooothoo  e
eotoho oo ooho e tho
ooho toho ehe  he t
ohe ho t he t to ooo
 t to tho ohehe to t
oooo t to t to ooo
he t t thoho tht tot
oht tho ho t ooo oo
ohhs the ththe t to
e toho hs e t tooto
ho oot e the too too
 ootoohshe to e to o
eheootho tht eo to t
```

epoch 100

```
Aln rlls ron otls ha
Thens herllls anD al
Tha  tnthe  an Irel
ID ln iorlI WoTWen p
The srle inlWane he
The srls ahes ol erl
D alBI, than phln ra
Aeisron rln Bal ton
TherPs erl D a llnsr
An  tlls alllo ronis
The Wol is he  anerl
Ae Drll D Dhen sre
A toilBli Iherhe han
Thal ItetlI han hll
Ihlres oIWeie i thel
BIi ID tnIt IhItha P
In anBhe pTle  tle p
Bln pItll  Jln Than
```

epoch 200

```
pThen the pa,r herso
Ae  te popto phas on
Thers aerpoaro pare
Ae parehere porppat
Appp,p Ae, Aar phere
Therppher hen,s ahrs
A, ate  a, pars aort
nhate po,ppt, thereo
Tho ptnpen  her paer
,o  atrpo pa,s anrsp
pnto tptort pheretrs
To, porpéptoerptater
Aarspa, po, pn,éos e
Aot ptrts herpnterpo
,hptteppt po toa en
Aere po, panos onrpa
Téts tn pe, pate pha
p AersTae onronrenes
```

epoch 300

```
The  pures ou Jo, ,
Bht e  fet moroumo,
I Aon When tnátman p
AenImotnáraul,  toel
Tre ronin thete, Jhe
Iu aes fatfIt to  ha
Wfumoe arers éerlára
Aen  anepf the mosco
Sfs hes Was of con,
Copoil tharl Wfe,  o
Heroa oaln the fare
In aneren  rf the fr
Het ateIá raul,  are
I Won  ie  au Hen, m
I  he motoals oae  a
fot is raicDramlronl
Solna maule  aremsph
```

epoch 400

Experiences in my NLP/ML/DL research life

- Most NLP problems need to capture context.
- SOP is not your SOP.
- Before mastering deep neural network, do not forget to try methods that you feel they will definitely fail.
- Feature engineering is a poison to overfitting, but is also a must to find the truth.
- LeCun is my Hero. (Though I also think machine learning is akin to “alchemy”.)

Thanks some of my students from IKM Lab.