Chinese Textual Sentiment Analysis: Datasets, Resources and Tools

Natural Language and Knowledge Processing Lab

Wei-Fan Chen and Lun-Wei Ku

December 11 @ Coling 2016, Osaka, Japan

Program and Speaker

Lecturer: Lun-Wei Ku

- 1. Overall Introduction (40 min)
- 2. Introduction to CSentiPackage (40 min)

-------------Coffee Break: 20 min -------------------

Lecturer: Wei-Fan Chen

- 3. Introduction to CSentiPackage:UTCNN (20 min)
- 4. Hands on Real data (40 min)

Overall Introduction

Sentiment Analysis

Sentiment Analysis Is…

- Studying opinions, sentiments, subjectivities, affects, emotions, views, etc. in text such as news, blogs, reviews, comments, dialogs, or other kind of documents.

- An important research question:
  - Sentiment information is global and powerful.
  - Sentiment information is valuable for companies, customers and personal communication.

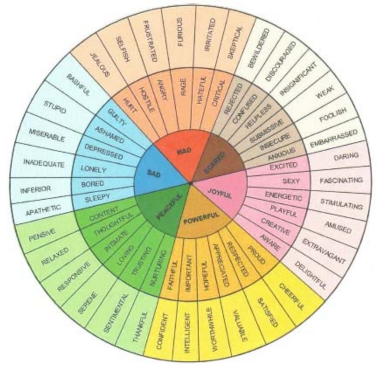
Opinion Definition

- From triple to quintuple
  - Triple:
    (e_j, so_{ij}, h_i)
  - Quintuple: (Bin Liu, NLP handbook, 2010)
    (e_j, a_{jk}, so_{ijkl}, h_i, t_l)

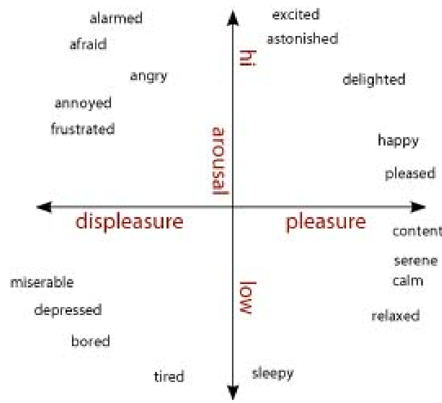
    e_j: target entity *j*
    h_i: holder *i*
    a_{jk}: aspect *k* (or sometimes called feature) of target entity *j*
    t_l: time *l*
    so: sentiment value of the opinion

Sentiment Representation

- Categorical
  - Sentiment, non-sentiment
  - Positive, neutral, negative
  - Stars
  - Emotions categories like Joy, Angry, Sadness…



- Dimensional
  - Valence Arousal



Sentiment Data Construction

- Sentiment labels are subjective: more annotators could make them more reliable.
- Manual gold data
  - Annotated by at least 3 annotators
  - Crowdsourcing
- User generated data (automatically generated)
  - User review scores (stars)
  - User generated text with emoticons (noisy)
  - Labels available from social platform

Annotation Consideration

- Granularity : Word, Sentence, Passage, Document?
  - Sentences are natural units but their labels are rarely found.
  - Detecting emotions from sentences is the most difficult (some are of complex semantic but very few words).
- Data Management
  - Explicit answer vs. majority answer
  - w/ context vs. w/o context
  - Data segmentation

Annotation Quality

- Agreement
  - Raw agreement
  - Kappa value, weighted kappa value

Now we get some ideas of sentiment analysis…let's see what the recent research is about!

Overall Introduction

Related Work

Widely known early work

- Thumbs up? Sentiment classification using machine learning techniques (Pang and Lee, EMNLP 2002): binary SVM classifier on documents.

A good start to get the idea of sentiment analysis

- Survey: Opinion Mining and Sentiment Analysis, Bo Pang and Lillian Lee, Foundations and Trends in Information Retrieval, 2008. (135 pages)

- Book: Sentiment Analysis and Opinion Mining, Bing Liu, Morgan & Claypool Publishers, 2012. (168 pages)

Recent One Year's Research… ACL

- Sentiment **Domain Adaptation** with Multiple Sources
- Connotation Frames: A **Data-Driven** Investigation
- Bi-Transferring **Deep Neural Networks** for **Domain Adaptation**
- Document-level **Sentiment Inference** with Social, Faction, and Discourse Context

Recent One Year's Research… NAACL

- Ultradense **Word Embeddings** by Orthogonal Transformation
- Separating **Actor-View** from **Speaker-View** Opinion Expressions using **Linguistic Features**
- Clustering for Simultaneous Extraction of **Aspects and Features** from Reviews
- Opinion Holder and Target Extraction on Opinion Compounds -- A **Linguistic Approach**
- Capturing **Reliable Fine-Grained Sentiment** Associations by **Crowdsourcing** and Best–Worst Scaling

Recent One Year's Research… EMNLP

- **Aspect** Level Sentiment Classification with **Deep Memory Network**
- Lifelong-RL: Lifelong Relaxation **Labeling** for Separating Entities and **Aspects** in Opinion Targets
- Learning **Sentence Embeddings** with Auxiliary Tasks for **Cross-Domain** Sentiment Classification
- **Attention-based LSTM Network** for **Cross-Lingual Sentiment** Classification

Recent One Year's Research…

- Aspect
- Domain Adaptation for Cross-Domain/Lingual
- Deep Neural Network vs. Linguistic Features
- Fine-Grained
- Crowdsourcing

Overall Introduction

Chinese Text Processing

Chinese Language

- Has no space between words
- The finest granularity of most sentiment tools is word : need word segmentation
- Part of speech tagging and syntactic information (parse tree) are nice to have.
- Two major Chinese writing forms: simplified Chinese and traditional Chinese

Chinese Language Processing Tools

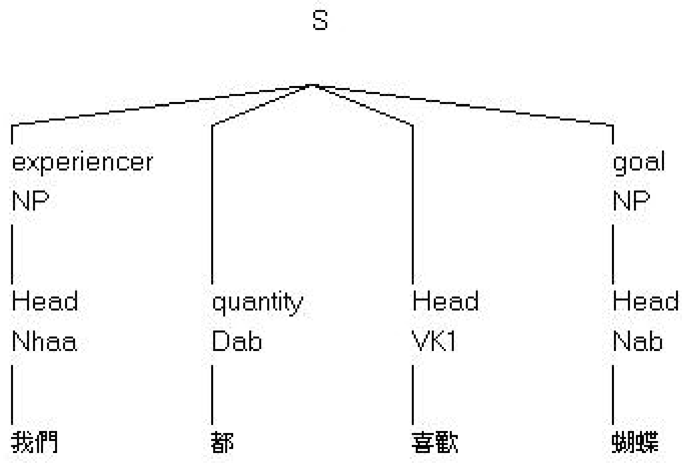
- The most widely used tool for Chinese is Stanford CoreNLP[1] (simplified Chinese)
- Other popular ones:
  - LTP Cloud (simplified Chinese)
  - CKIP Parser[2] (traditional Chinese)
  - jieba (segmentation, both simplified/traditional Chinese)

1 http://nlp.stanford.edu/software/

2 http://godel.iis.sinica.edu.tw/CKIP/parser.htm

CKIP Parser

- Its tag set is different from Stanford CoreNLP's



我們都喜歡蝴蝶

我們(Nh) 都(D) 喜歡(VK) 蝴蝶(Na)

#1:1.[0] S(experiencer:NP(Head:Nhaa:我們)|quantity:Dab:都|Head:VK1:喜歡|goal:NP(Head:Nab:蝴蝶))#。(PERIODCATEGORY)

- We provide a tag mapping file (for sentiment analysis)

CSentiPackage @NLPSA

CSentiPackage

- Datasets
  - Chinese Morphological Dataset Cmorph (former version of ACiBiMA)*
  - Chinese Opinion Treebank
- Resources
  - NTUSD/ANTUSD
- Tools
  - CopeOpi + Tag Mapping File
  - UTCNN

*https://github.com/windx0303/ACBiMA

Statistics

- NTUSD: Sentiment Dictionary (with 10,371 words): free for research, 400+ applications
- ANTUSD: Augmented NTUSD (with 27,221 words, now integrating with e-Hownet)
- Cmorph (with 8,000+ words) -> ACBiMA (with 11,000+ words)
- Chinese Opinion Treebank: labels on Chinese Treebank 5.1

Materials:
From Words to Sentences

- NTUSD: words (binary sentiment)
- ANTUSD: words (annotation features)
- Chinese Morphological Dataset: words (morphological structures)
- Chinese Opinion Treebank: phrases (sentence structure)
- Chinese Opinion Treebank: sentences (binary sentiment)

Tools:
From Words to Sentences, Documents, and Beyond

- CopeOpi Sentiment Scoring Tool: words, sentences, documents, documents+ (text)
- UTCNN: posts and users (text and social media)

NTUSD

- Simplified Chinese and traditional Chinese versions
- A positive word collection of 2,812 words
- A negative word collection of 8,276 words
- No degree, no estimated scores and other information.

ANTUSD

- 6 Fields
  - CopeOpi Score
  - Number of positive annotation
  - Number of neutral annotation
  - Number of negative annotation
  - Number of non-sentiment annotation
  - Number of not-a-word annotation

開心	0.434168	1	0	0	0	0
酣聲	0	0	0	1	3	0
憤怒	-0.80011	0	0	5	0	0

- Not-a-word: useful as they are collected from real segmentated data

ANTUSD

- Contains also short phrases like一昧要求, 一路過關斬將,備受外界期待…

ANTUSD and E-HOWNET

E-HowNet

- A frame-based entity-relation model extended from HowNet
- Define lexical senses (concepts) in a hierarchical manner
- Now integrated with ANTUSD and covers 47.7% words in ANTUSD

- An integration of two resources which may help us play with sentiment and semantics.
- Related English resource: SentiWordnet
  - Refer to Wordnet
  - With PosScore and NegScore added
  - ObjScore = 1-(PosScore+NegScore)

ANTUSD in E-HOWNET

詞彙:	致勝	Word
詞性:	VH11	Pos Tag
英文意涵:	win victory	English Meaning
概念式:	{win|獲勝}	Concept Frame
展開式:		
WordNet 自動連結:	{gain.v.05, succeed.v.01, acquire.v.05, win.v.01}	WordNet Linkage

Sentiment					
score	**positive**	**neutral**	**negative**	**non_opinion**	**non_word**
0.5772	1	0	0	0	0

ExistAppear|存現

disappear|消失 [一掃而空，不見了，不知去向，不翼而飛，化為烏有，幻滅，石沉大海，冰消瓦解，沒，杳如黃鶴，杳無信息，杳無音訊，
消，消失，消退，消逝，消逝無蹤，消散，消褪，消聲匿跡，破空，破滅，退去，脫漏，逝，逝去，逐波而去，散佚，渙，絕跡，雲消霧散，
，隱沒，飄逝，瓔滅]

BeNormal|常態

BeRecovered|復原 [平復，息事，復元，復原，復甦，穌，還原]

BeGood|良態

BeFull|吃飽 [吃飽，吃飽喝足，酒足飯飽，飫，飽足，飽脹，鼓腹，饜]

lucky|幸運 [三生有幸，平順，吉人天相，行大運，走好運，走運，事事如意，和氣致祥，時運亨通，桃花運，泰順，得時，開泰，順常，
僥倖，儌，徼幸，邀天之幸，雙喜臨門]

prosper|發達 [方興未艾，水漲船高，功成名就，功成名遂，平步青雲，未艾方興，亨，亨通，壯盛，走高，昌盛，爭氣，長進，勃發，
入室，登龍，進化，進步，進展，新高，鼎盛，蒸蒸日上，繁榮，鯉躍龍門，鵬程萬里，騰達，出頭，顯耀]

win|獲勝 [打勝，打勝仗，告捷，取勝，拔尖，奏捷，致勝，得勝，脫穎而出，凱，勝，勝利，獨占鰲頭，獨占鰲頭，獨佔鰲頭，獨佔鰲

獲選|BeSelected [當選，獲選，膺選]

OtherWord(win|獲勝)

surpass|強過 [以小吃大，占上風，有過之無不及，佔上風，青
當先，趕過，遙遙領先，領先，獨步，優於，壓倒，賽，蓋，

WellKnown|成名 [人口皆碑，出名，光宗耀祖，成名，有口皆
]

succeed|成功 [大功告成，出人頭地，成功，成事，收效，有

able|能 [力所能及，又紅又專，允文允武，文武全才，文武合一
科班出身，拿手，神通廣大，純熟，能文能武，能幹，高桿，專
嫻熟，熟巧，熟妙，熟習，熟練，熟爛，練達，駕輕就熟，諳

BeBad|衰變

end|終結 [中止，止息，休止，告終，終止，終結，斷，止]

WeatherState|天候狀態

MentalState|精神狀態

MentalAct|精神動作

change|變化 [化，幻化，日新月異，生變，白雲蒼狗，改樣，改觀，
，變出，變動，變遷，變易，起落]

corelation|關連

AttributeValue|屬性值

act|行動 [行動]

OtherWord(event|事件)

object|物體 [事物，客體，對象]

relation

詞彙訊息

詞彙:	致勝
詞性:	VH11
英文意涵:	win victory
Event Frame:	
定義式:	{win|獲勝}
操作式:	
語義功能:	
語義特徵:	
展開式:	
WordNet 自動連結:	{gain.v.05, succeed.v.01, acquire.v.05, win.v.01}

Sentiment

score	positive	neutral	negative	non_opinion	non_word
0.5772	1	0	0	0	0

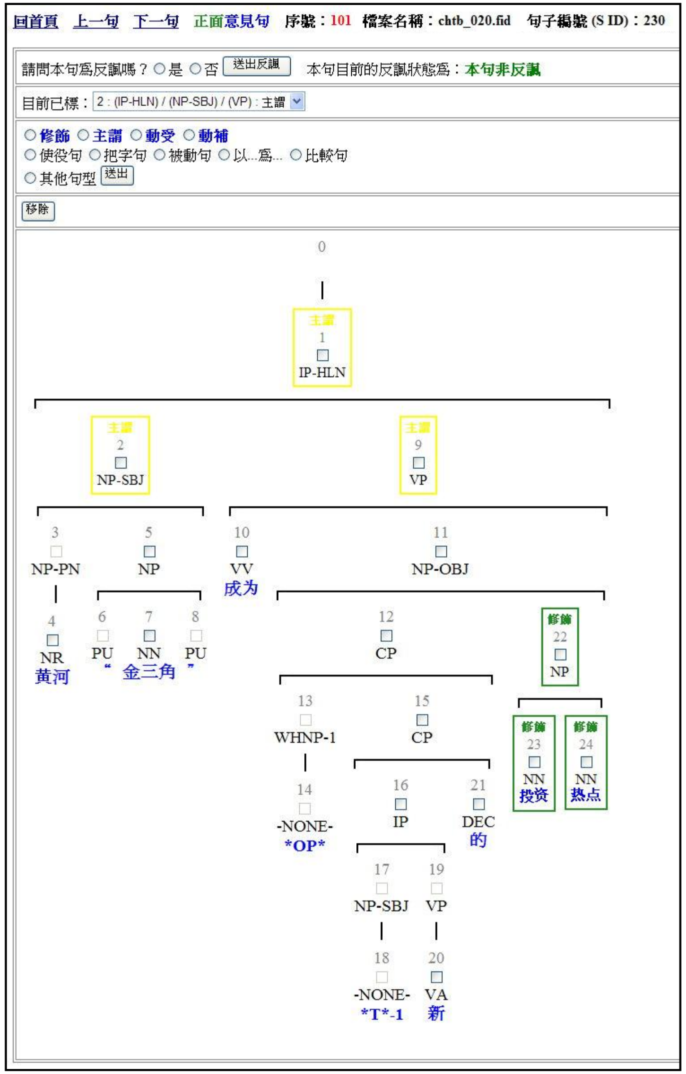
Chinese Morphological Structure

- Parallel type: 財富 (rich wealth)
- Substantive-Modifier type: 痛哭 (bitterly cry)
- Subjective-Predicate type: 山崩 (land slip; landslide)
- Verb-Object type: 避暑 (escape from summer)
- Verb-Complement type: 提高 (increase: raise up)
- Negation type: 無情 (no feelings)
- Confirmation type: 有心 (have heart)
- Others

Chinese Opinion Treebank

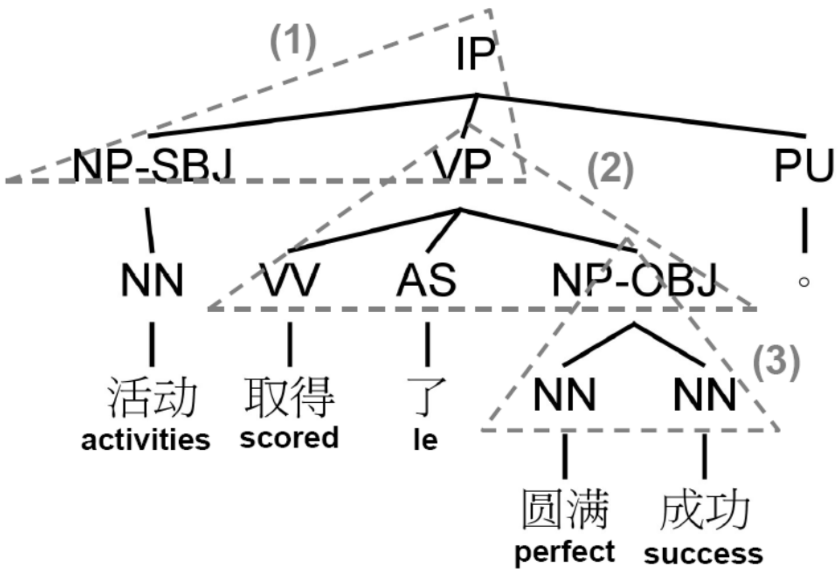
- Based on Chinese Treebank 5.1.
- Including the opinion labels of each sentences.
- Including the word-pairs and their composing type in opinionated sentences.
- To avoid copyright issue, you need to have Chinese Treebank 5.1 by yourself in order to use Chinese Opinion Treebank!

Chinese Opinion Treebank



S ID=230: 黄河“金三角”成为新的投资热点		
.node file	.tree file	.trio file
Fields		
Node ID, POS, node content, node depth	Node ID: children	Trio ID, trio head, trio left node, trio right node, trio type
Content		
0,,,0	0:1,	2,1,2,9,3
1,IP-HLN,,1	1:2,9,	3,22,23,24,2
2,NP-SBJ,,2	2:3,5,	
3,NP-PN,,3	3:4,	
4,NR,黄河,4	4:	
5,NP,,3	5:6,7,8,	
6,PU,“,4	6:	
7,NN,金三角,4	7:	
8,PU,”,4	8:	
9,VP,,2	9:10,11,	
10,VV,成为,3	10:	
11,NP-OBJ,,3	11:12,22,	
12,CP,,4	12:13,15,	
13,WHNP-1,,5	13:14,	
14,-NONE-,*OP*,6	14:	
15,CP,,5	15:16,21,	
16,IP,,6	16:17,19,	
17,NP-SBJ,,7	17:18,	
18,-NONE-,*T*-1,8	18:	
19,VP,,7	19:20,	
20,VA,新,8	20:	
21,DEC,的,6	21:	
22,NP,,4	22:23,24,	
23,NN,投资,5	23:	
24,NN,热点,5	24:	
Opinion labels of three annotators (filename, SID, opinion, polarity, opinion type)		
chtb_020.raw,230,N,, chtb_020.raw,230,Y,POS,STATUS chtb_020.raw,230,Y,POS,STATUS		
Opinion gold standard		
chtb_020.raw,230,Y,POS,STATUS		

Notation (Parsing Tree)



Tri(S)=
1, IP, 活动, VP, Subjective-Predicate
2, VP, 取得, NP-OBJ, Verb-Object
3,NP-OBJ, 圆满, 成功, Substantive-Modifier

- *T*: the parsing tree of a sentence *S*
- $O = \{o_1, o_2, \ldots\}$: in-ordered set of tree nodes
- *tri* = $(triID, o_{parent}, o_{left}, o_{right}, t) \in Tri$: an opinion trio
- $t \in Rpt$: a syntactic inter-word relation
  Rpt ϵ {*Substantive-Modifier*, *Subjective-Predicate*, *Verb-Object*, *Verb-Complement*, *Other*}

Chinese Opinion Treebank

- Align the opinion labels of sentences to Chinese Treebank 5.1 by sentence IDs.
- Align Opinion trios to Chinese Treebank 5.1 by node IDs.
- Can be used to do opinion cause analysis.

CopeOpi

- A statistical sentiment analysis tool
- Can be used without any training
- Users can update character weights or add any sentiment words
- It runs fast.

The First Idea

- Chinese characters are mostly morphemes and they bear sentiment, too.
- Simple example: some characters are preferred for naming, but some are not.
- For example, 德(ethic) 胜(win) 高(high) good for names; 笨(stupid) 悲(sorrow) 惨(terrible) are not good choices for names.
- With some exceptions, but still quite reliable if the sentiment of character is acquired statistically from a large naming corpus (or just sentiment dictionaries.) Exceptions like 徐悲鸿.

Bag of Unit

$$N_{c_i} = \frac{fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \qquad P_{c_i} = \frac{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}}$$

$$S_{c_i} = (P_{c_i} - N_{c_i})$$

$$S_w = \frac{1}{p} \times \sum_{j=1}^{p} S_{c_j}$$

[仇 (-1.0) + 視 (0.0)] / 2 = -1/2 = -0.5 (NEG)
[富(1.0) + 貴(0.936)] / 2 = 0.968 (POS)

好人、美麗、憤怒、弱小…

Aggregation

- Word sentiment
  - Summing up opinion scores of characters
- Sentence sentiment
  - Summing up opinion scores of words

So is there any way we can give them weights?

Weighted by Structures

- Linguistic Information:
  - Morphological structures
    - Intra-word structures
  - Sentence syntactic structures
    - Inter-word structures

Morphological Structure

Get types by SVM, CRF, handcraft…

Linguistic Morpho. Type	Example
1. Parallel	財富、打罵
2. Substantive-Modifier	低級、痛哭
3. Subjective-Predicate	心疼、氣虛
4. Verb-Object	失控、免職
5. Verb-Complement	看清、擊潰
Opinion Morpho. Type	**Example**
6. Negation	無法、不慎
7. Confirmation	有賴、有愧
8. Others	姪子、薄荷

Example of Sentiment Trios in Chinese Opinion Treebank

Linguistic Morpho. Type	Example
Parallel (Skip)	美麗而聰慧
1. Substantive-Modifier	高大的樓房
2. Subjective-Predicate	學習認真
3. Verb-Object	恢復疲勞
4. Verb-Complement	收拾乾淨
Morpho. Type Opinion	Example
n. Others	為…/以…

Compositional Chinese Sentiment Analysis

Sentiment Scoring Formula for Each Morphological Type:

- Parallel type

$$S(C_1C_2) = \frac{S(C_1) + S(C_2)}{2}$$

- Substantive-Modifier type

$$
\begin{aligned}
&\text{if } (S(C_1) \neq 0 \text{ and } S(C_2) \neq 0) \text{ then} \\
&\quad \text{if } (S(C_1) > 0 \text{ and } S(C_2) > 0) \text{ then } S(C_1C_2) = S(C_1) \\
&\quad \text{else } S(C_1C_2) = -1 \times |S(C_1)| \\
&\text{else } S(C_1C_2) = S(C_1) + S(C_2)
\end{aligned}
$$

- Subjective-Predicate

$$
\begin{aligned}
&\text{if } (S(C_2) \neq 0) \text{ then } S(C_1C_2) = S(C_2) \\
&\text{else } S(C_1C_2) = S(C_1)
\end{aligned}
$$

- Example:氣虛
- Subjective-Predicate type
- 氣 0.5195
- 虛 -0.8178
- Score(氣虛) = -0.8178

Compositional Chinese Sentiment Analysis

Sentiment Scoring Formula for Each Morphological Type:

- Example:看清、看壞
- Verb-Complement type
- 看: 0.1
- 清: 0.8032
- 壞: -0.9
- Score(看清) = 0.8072
- Score(看壞) = -0.9

- Verb-Object type

$$\begin{aligned}&\text{if } (S(C_1) \neq 0 \text{ and } S(C_2) \neq 0)\\&\quad \text{then } S(C_1C_2) = |S(C_1)| \times SIGN(S(C_1)) \times SIGN(S(C_2))\\&\quad \text{else } S(C_1C_2) = S(C_1) + S(C_2)\end{aligned}$$

- Verb-Complement type = Subjective-Predicate type
- Negation type

$$\begin{aligned}&\text{if } (C_1 \in NC) \text{ then } S(C_1C_2) = (-1) \times S(C_2)\\&\text{else } S(C_1C_2) = (-1) \times S(C_1)\end{aligned}$$

- Confirmation type

$$\text{if } (C_1 \in PC) \text{ then} S(C_1C_2) = S(C_2) \text{ else} S(C_1C_2) = S(C_1)$$

- Others = Parallel type

Example of Using Sentiment Trios

- Score: 0.6736

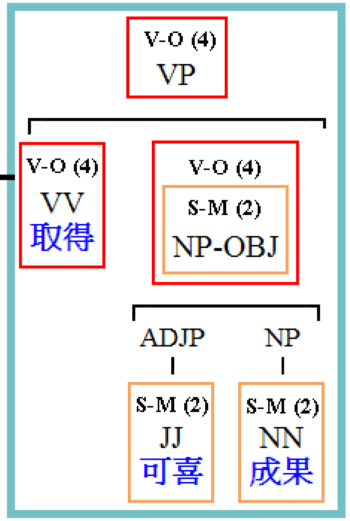
⇧

0.3018

Verb-Object type

0.6736

$$\text{if } (S(C_1) \neq 0 \text{ and } S(C_2) \neq 0)$$
$$\text{then } S(C_1C_2) = |S(C_1)| \times SIGN(S(C_1)) \times SIGN(S(C_2))$$
$$\text{else } S(C_1C_2) = S(C_1) + S(C_2)$$



Substantive-Modifier type

$$\text{if } (S(C_1) \neq 0 \text{ and } S(C_2) \neq 0) \text{ then}$$
$$\quad \text{if } (S(C_1) > 0 \text{ and } S(C_2) > 0) \text{ then } S(C_1C_2) = S(C_1)$$
$$\quad \text{else } S(C_1C_2) = -1 \times |S(C_1)|$$
$$\text{else } S(C_1C_2) = S(C_1) + S(C_2)$$

0.6736

0.4109

Performance of CopeOpi (Dataset w/o Structure)

Level	Corpus	By	Precision	Recall	f-measure
Word	836 words	Annotator	0.81	0.80	0.80
Sentence	CIRB010-OP	Annotator	0.75	0.65	0.66
Document	CIRB010-OP	Annotator	0.73	0.69	0.72
Word	836 words	Machine	0.61	0.79	0.68
Sentence	CIRB010-OP	Machine	0.38	0.65	0.48
Sentence	CIRB020-OP	Machine	0.33	0.45	0.38
Sentence	CIRB020-OP-R	Machine	0.66	0.89	0.76
Document	CIRB010-OP	Machine	0.40	0.55	0.46

*NTCIR MOAT Corpus as materials

Performance of CopeOpi (Dataset w/ Structure)

Setting	Word	Sentence	Opinion	Polarity	Desc
1	bag	bag	0.7073	0.4988	
2	struc	bag	0.7162	0.5117	CRF
3	bag	struc	0.8000	0.5361	Manual
4	struc	struc	0.7922	0.5297	CRF+Manual
5	struc	struc	0.7993	0.5187	CRF+Auto

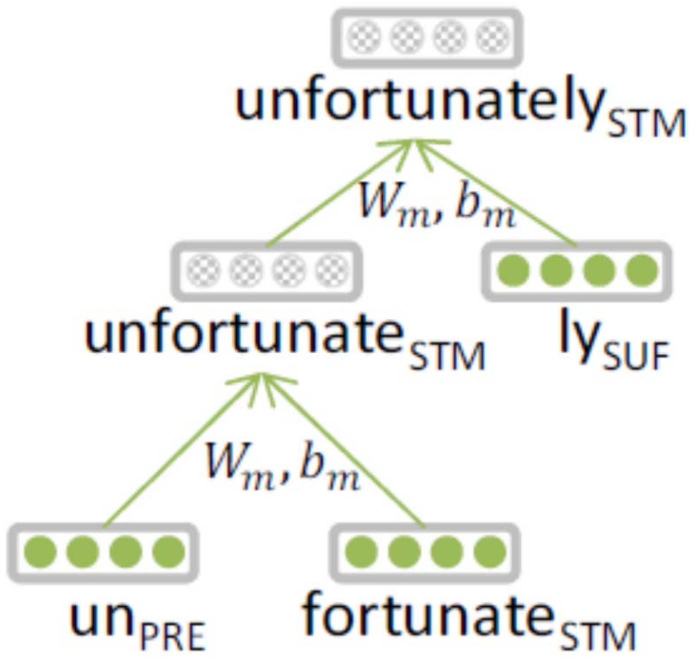
*Chinese Opinion Treebank as materials

Performance of CopeOpi (FB Stance Classification)

Method	Precision			Recall			F-score			F_1^{SNU}
	Sup	Neu	Uns	Sup	Neu	Uns	Sup	Neu	Uns	
Majority	.000	.908	.000	.000	1.00	.000	.000	.952	.000	.317 (-39%)
Graph-joint	.564	.958	.000	.631	.955	.000	.596	.956	.000	.518 (—)
Dic	.102	.929	.007	.066	.148	.773	.080	.255	.014	.337 (-35%)
Graph-sentiment	.550	.999	.000	.992	.932	.000	.707	.965	.000	.572 (+10%)
Graph-joint	.564	.958	.000	.631	.955	.000	.596	.956	.000	.518 (—)
SVM-Uni+Bi+TriGram	.470	.918	1.00	.121	.988	.045	.192	.952	.087	.519 (+0%)
No engagement (CopeOpi)	.596	.971	.056	.198	.970	.500	.297	.970	.101	.548 (+6%)
Joint-MRF	.697	.971	.400	.724	.970	.273	**.710**	.970	**.324**	**.672 (+30%)***
Random cold start	.086	.912	.032	.673	.329	.045	.152	.483	.038	.346 (—)
SVM cold start	1.00	.910	.000	.019	1.00	.000	.038	.953	.000	.443 (+28%)
SVM cold start	1.00	.910	.000	.019	1.00	.000	.038	.953	.000	.443 (+28%)

Deep Neural Network Example Word

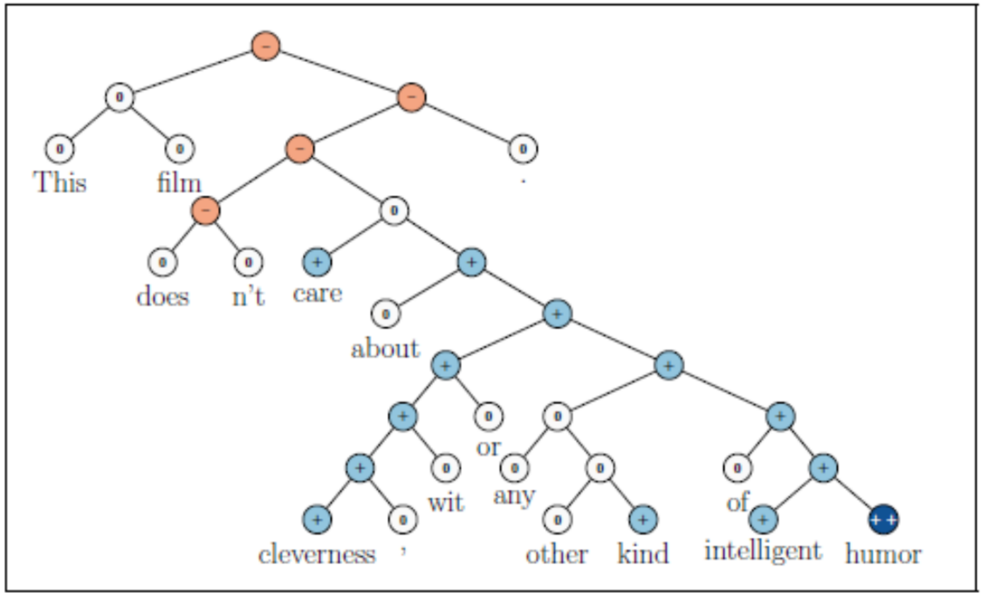
- Morphological structure for a better word representation.
- Same idea but for *Chinese sentiment analysis*

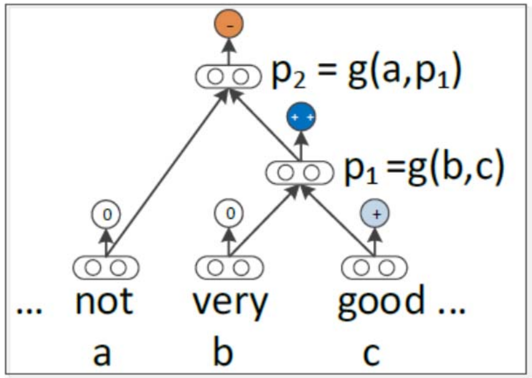


- Luong, Thang, Richard Socher, and Christopher D. Manning. "Better Word Representations with Recursive Neural Networks for Morphology." *CoNLL*. 2013.

Deep Neural Network Example Sentence

- Learned composition function (of semantics): Richard Socher (RNN, series work from 2011)





Learning by Neural Network

- Word Sentiment
- Sentence Sentiment
- Document Sentiment
- Social Media Post Sentiment

Learning by Deep Neural Network

- **Word Sentiment: CNN + ANTUSD**
- Sentence Sentiment
- Document Sentiment
- **Social Media Post Sentiment: Text + User Context**

  – **Not yet consider structures!**

Word Sentiment NN: CNN + ANTUSD

A Demonstrative Experiment

ANTUSD: A Large Chinese Sentiment Dictionary, Shih-Ming Wang and Lun-Wei Ku, in Proceedings of LREC 2016

Experiment Setting

- Dataset: ANTUSD ∩ E-hownet, a total 12995 words
- Classifier: support vector machine (SVM) with linear kernel
- Average over 10-fold validation scores

Three sentiment analysis tasks

- Opinion extraction: identify opinion words ({**POS**,**NEG**} v.s. **NONOP**)
- Polarity classification: classify opinion words (**POS** v.s. **NEG**)
- Combined tasks (**POS**, **NEG**, **NONOP**)
  - $P = \frac{correct(opinion) \cap correct(polarity)}{proposed(opinion)}$
  - $R = \frac{correct(opinion) \cap correct(polarity)}{gold(opinion)}$
  - $F\_score = \frac{2PR}{P+R}$

Preprocessing

Extract single label for each word

1. **NOT**: Count(Not)>0
2. **NONOP**: Count(Non)>0
3. **POS**: Count(Pos)>0 and Count(Neg)=0
4. **NEG**: Count(Neg)>0 and Count(Pos)=0
5. **NEU**: Count(Pos)=0, Count(Neg)=0 and Count(Neu)>0

Preprocessing

Extract single label for each word

1. **NOT**: Count(Not)>0
2. **NONOP**: Count(Non)>0
3. **POS**: Count(Pos)>0 and Count(Neg)=0
4. **NEG**: Count(Neg)>0 and Count(Pos)=0
5. **NEU**: Count(Pos)=0, Count(Neg)=0 and Count(Neu)>0

- **NOT** words are not used
- **NEU** words are dropped since there are only 16 of them

Features

ANTUSD & E-hownet

- CopeOpi score in ANTUSD
- Synonym-Set index (SSI)
  - Concept frame index of a word
  - Each word might belong to many concepts
  - Represented as a binary vector

Features

ANTUSD & E-hownet

- CopeOpi score in ANTUSD
- Synonym-Set index (SSI)
  - Concept frame index of a word
  - Each word might belong to many concepts
  - Represented as a binary vector

Word Embedding

- Corpus: LDC2009T14 (Chinese news)
- Word vectors
- Summation of char vectors
  - Very high coverage rate

Opinion Extraction

- COP, SSI has lower precision
  - opinion extraction is more semantic-oriented
  - Many concept frame contain only one word

Feature(s)	Precision	Recall	f-score
COP	0.686	1.000	0.814
SSI	0.693	0.993	0.816
WV	0.784	0.936	0.854
CV	0.765	0.919	0.835
COP+SSI	0.740	0.914	0.818
COP+WV	0.785	0.933	0.853
COP+CV	0.764	0.917	0.833
SSI+WV	0.789	0.937	0.856
SSI+CV	0.772	0.920	0.840
WV+CV	0.808	0.921	0.861

Opinion Extraction

- COP, SSI has lower precision
  - opinion extraction is more semantic-oriented
  - Many concept frame contain only one word
- Character vectors lead to slightly worse performance

Feature(s)	Precision	Recall	f-score
COP	0.686	1.000	0.814
SSI	0.693	0.993	0.816
WV	0.784	0.936	0.854
CV	0.765	0.919	0.835
COP+SSI	0.740	0.914	0.818
COP+WV	0.785	0.933	0.853
COP+CV	0.764	0.917	0.833
SSI+WV	0.789	0.937	0.856
SSI+CV	0.772	0.920	0.840
WV+CV	0.808	0.921	0.861

Opinion Extraction

- .., COP, SSI has lower precision
  - .., opinion extraction is more semantic-oriented
  - .., Many concept frame contain only one word
- .., Character vectors lead to slightly worse performance
- .., Features are complemented; combined features leads to improvement

Feature(s)	Precision	Recall	f-score
COP	0.686	1.000	0.814
SSI	0.693	0.993	0.816
WV	0.784	0.936	0.854
CV	0.765	0.919	0.835
COP+SSI	0.740	0.914	0.818
COP+WV	0.785	0.933	0.853
COP+CV	0.764	0.917	0.833
SSI+WV	0.789	0.937	0.856
SSI+CV	0.772	0.920	0.840
WV+CV	0.808	0.921	0.861

Polarity Classification

- COP leads to a significant better result, reflecting is sentiment-oriented nature

Feature(s)	POS f1	NEG f1	Average f1
COP	0.973	0.976	0.974
SSI	0.792	0.842	0.817
WV	0.870	0.895	0.882
CV	0.829	0.851	0.840
COP+SSI	0.979	0.982	0.980
COP+WV	0.981	0.984	0.982
COP+CV	0.967	0.972	0.970
SSI+WV	0.898	0.915	0.907
SSI+CV	0.868	0.886	0.877
WV+CV	0.899	0.916	0.908

Polarity Classification

- COP leads to a significant better result, reflecting is sentiment-oriented nature
- Combining COP & other features still leads to improvement

Feature(s)	POS f1	NEG f1	Average f1
COP	0.973	0.976	0.974
SSI	0.792	0.842	0.817
WV	0.870	0.895	0.882
CV	0.829	0.851	0.840
COP+SSI	0.979	0.982	0.980
COP+WV	0.981	0.984	0.982
COP+CV	0.967	0.972	0.970
SSI+WV	0.898	0.915	0.907
SSI+CV	0.868	0.886	0.877
WV+CV	0.899	0.916	0.908

Polarity Classification

- COP leads to a significant better result, reflecting is sentiment-oriented nature
- Combining COP & other features still leads to improvement
- Combining word vectors and SSI also leads to improvement

Feature(s)	POS f1	NEG f1	Average f1
COP	0.973	0.976	0.974
SSI	0.792	0.842	0.817
WV	0.870	0.895	0.882
CV	0.829	0.851	0.840
COP+SSI	0.979	0.982	0.980
COP+WV	0.981	0.984	0.982
COP+CV	0.967	0.972	0.970
SSI+WV	0.898	0.915	0.907
SSI+CV	0.868	0.886	0.877
WV+CV	0.899	0.916	0.908

Combined Task

- COP outperforms the others

Feature(s)	Precision	Recall	f-score
COP	0.912	0.927	0.920
SSI	0.706	0.679	0.692
WV	0.737	0.767	0.752
CV	0.689	0.721	0.705
COP+SSI	0.864	0.945	0.903
COP+WV	0.850	0.902	0.875
COP+CV	0.840	0.869	0.854
SSI+WV	0.764	0.796	0.779
SSI+CV	0.732	0.755	0.743
WV+CV	0.764	0.813	0.787

Combined Task

- COP outperforms the others
- Both the numerator of precision and recall are affected by COP's better polarity classification ability
- Only the denominator of precision is affected by COP's worse opinion extraction ability

Precision & Recal l

$$P = \frac{correct(opinion) \cap correct(polarity)}{proposed(opinion)}$$

$$R = \frac{correct(opinion) \cap correct(polarity)}{gold(opinion)}$$

Feature(s)	Precision	Recall	f-score
COP	0.912	0.927	0.920
SSI	0.706	0.679	0.692
WV	0.737	0.767	0.752
CV	0.689	0.721	0.705
COP+SSI	0.864	0.945	0.903
COP+WV	0.850	0.902	0.875
COP+CV	0.840	0.869	0.854
SSI+WV	0.764	0.796	0.779
SSI+CV	0.732	0.755	0.743
WV+CV	0.764	0.813	0.787

Combined Task

- COP outperforms the others
- Both the numerator of precision and recall are affected by COP's better polarity classification ability
- Only the denominator of precision is affected by COP's worse opinion

  extraction ability
- WV+CV outperforms WV due to coverage issue

Feature(s)	Precision	Recall	f-score
COP	0.912	0.927	0.920
SSI	0.706	0.679	0.692
WV	0.737	0.767	0.752
CV	0.689	0.721	0.705
COP+SSI	0.864	0.945	0.903
COP+WV	0.850	0.902	0.875
COP+CV	0.840	0.869	0.854
SSI+WV	0.764	0.796	0.779
SSI+CV	0.732	0.755	0.743
WV+CV	0.764	0.813	0.787

Inject More Semantics: ANTUSD and E-Hownet

E-HowNet

- A frame-based entity-relation model extended from HowNet
- Define lexical senses (concepts) in a hierarchical manner
- Now integrated with ANTUSD and covers 47.7% words in ANTUSD

Wrapup

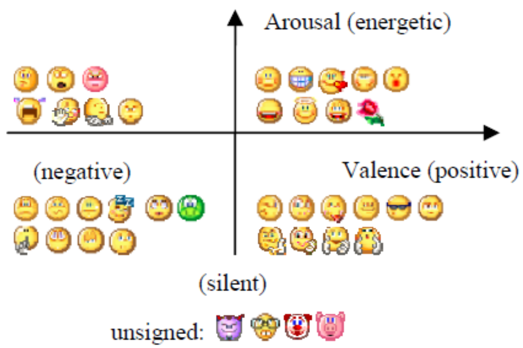
- CSentiPackage
  - NTUSD/ANTUSD/ANTUSD+e-HowNet
  - Chinese Morphological Dataset Cmorph
  - Chinese Opinion Treebank
  - CopeOpi + Tag Mapping File
  - An demonstrative exp of ANTUSD

  ========= We are here =================

  - UTCNN (next session)
- Hand-on

Future Release Tool in CSentiPackage

- EmotionPushCore: short message emotion detector (ongoing)




Coffee Break
shutterstock
IMAGE ID: 62982823
www.shutterstock.com

10:20-10:40

CSentiPackage: UTCNN

Learning by Deep Neural Network

- Word Sentiment: CNN + ANTUSD
- Sentence Sentiment
- Document Sentiment
- **Social Media Post Sentiment: Text + User Context**

Outline

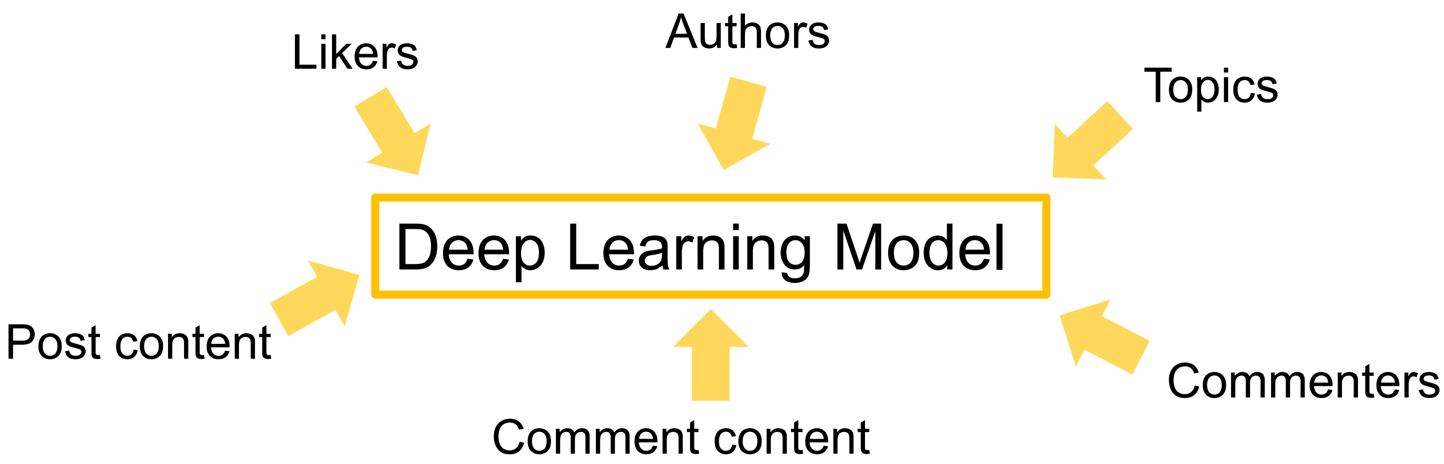
- CSentiPackage: UTCNN
  - Introduction
  - Model
  - Results
- Hands on real data
  - Environment
  - Data preprocessing
  - Tools
    - NTUSD and ANTUSD
    - Cmorph and Chinese Opinion Treebank
    - CopeOpi
    - UTCNN

Outline

- CSentiPackage: UTCNN
  - Introduction
  - Model
  - Results
- Hands on real data
  - Environment
  - Data preprocessing
  - Tools
    - NTUSD and ANTUSD
    - Cmorph and Chinese Opinion Treebank
    - CopeOpi
    - UTCNN

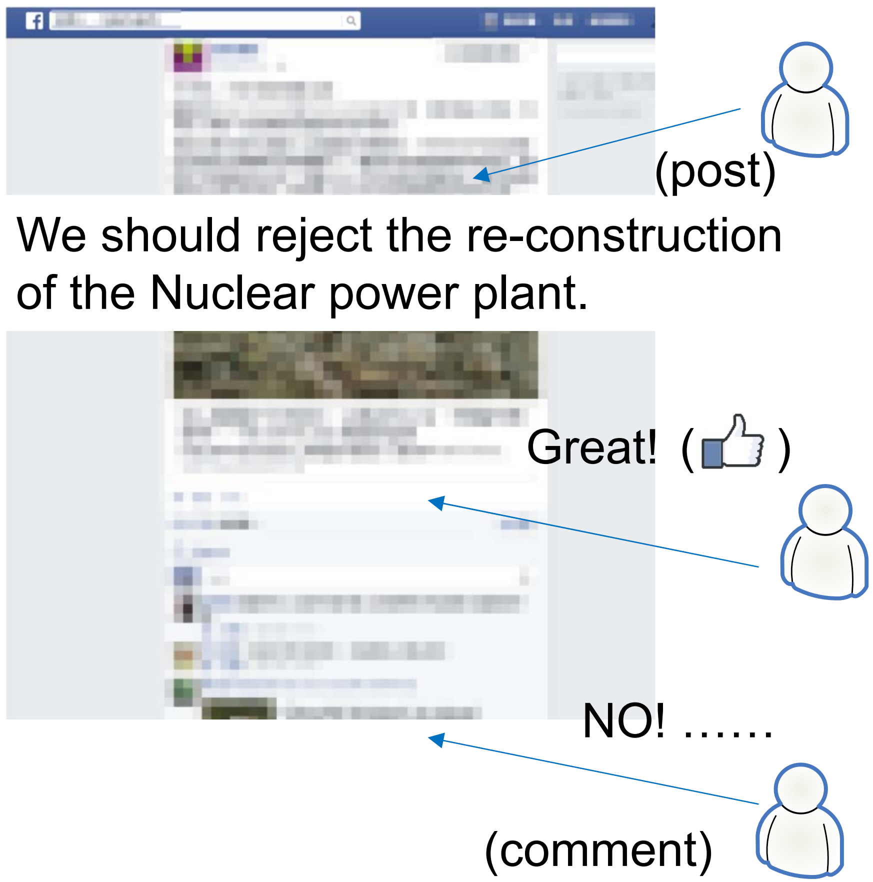
User Topic Comment Neural Network (UTCNN)

- A deep learning model of stance classification on social media text



UTCNN

- Stance tendency
  - Author
  - Liker
  - Topic
  - Commenter
- Semantic preference
  - Author
  - Liker
  - Topic
  - Commenter

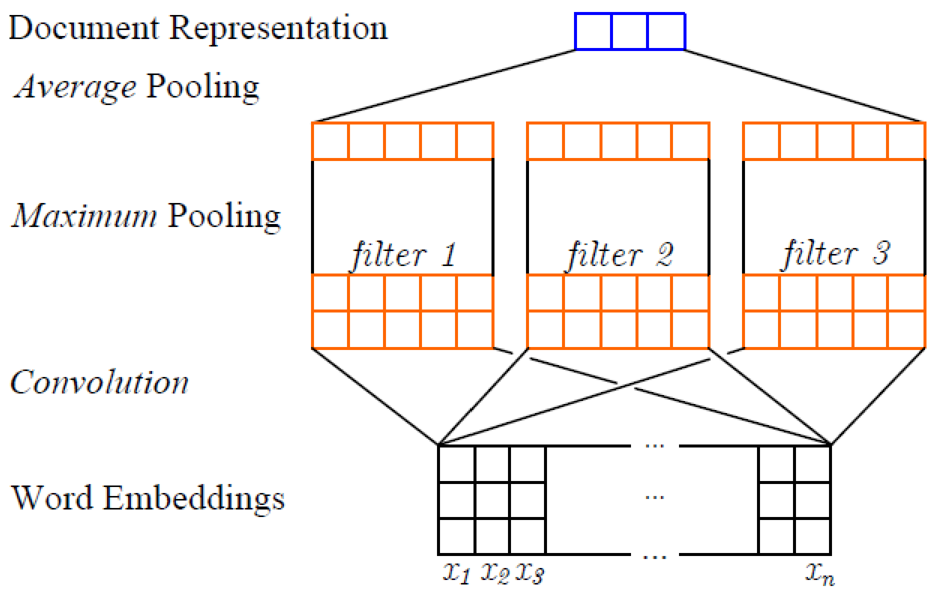


Document Composition

- From word representation to document representation
  - CNN
  - RNN
    - LSTM

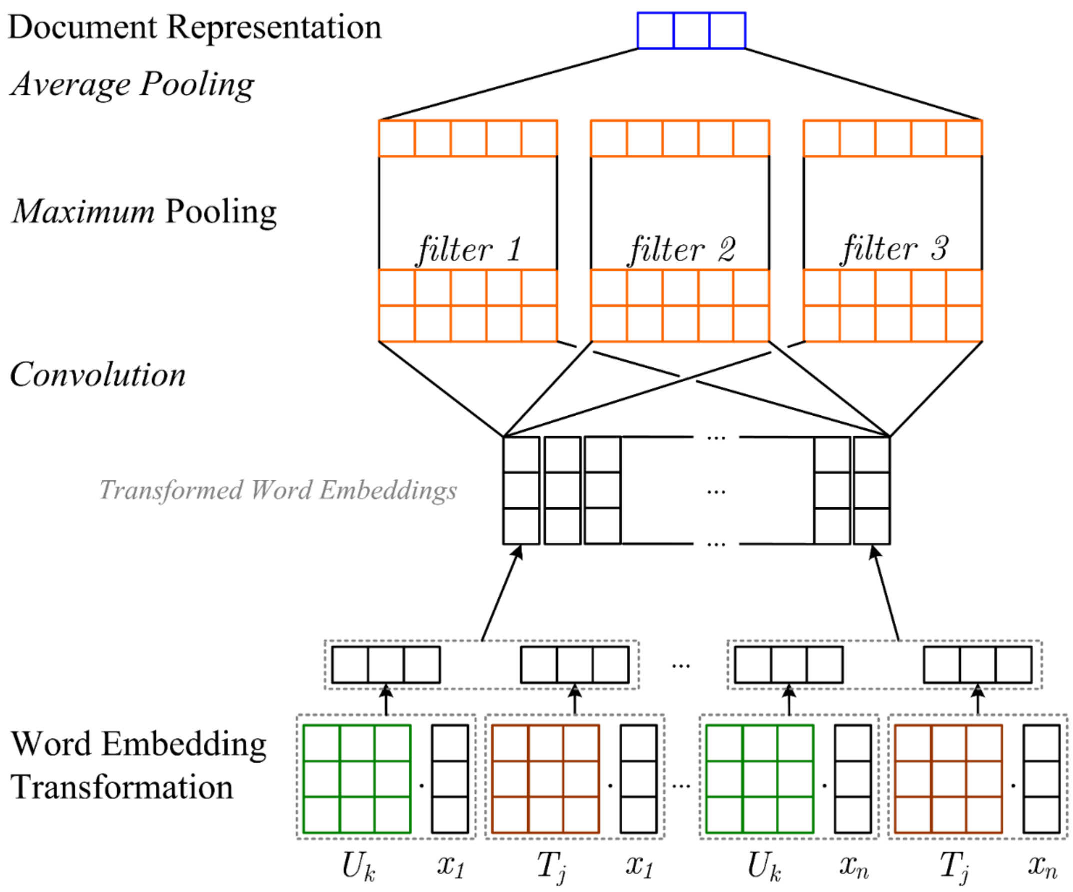
CNN architecture

- $x_c = [x_1; x_2; ...; x_n]$
- $h_{cf} = f\left(W_{cf} \cdot x_c + b_{cf}\right)$
- Capture *n*-gram features



User- and Topic-dependent document composition

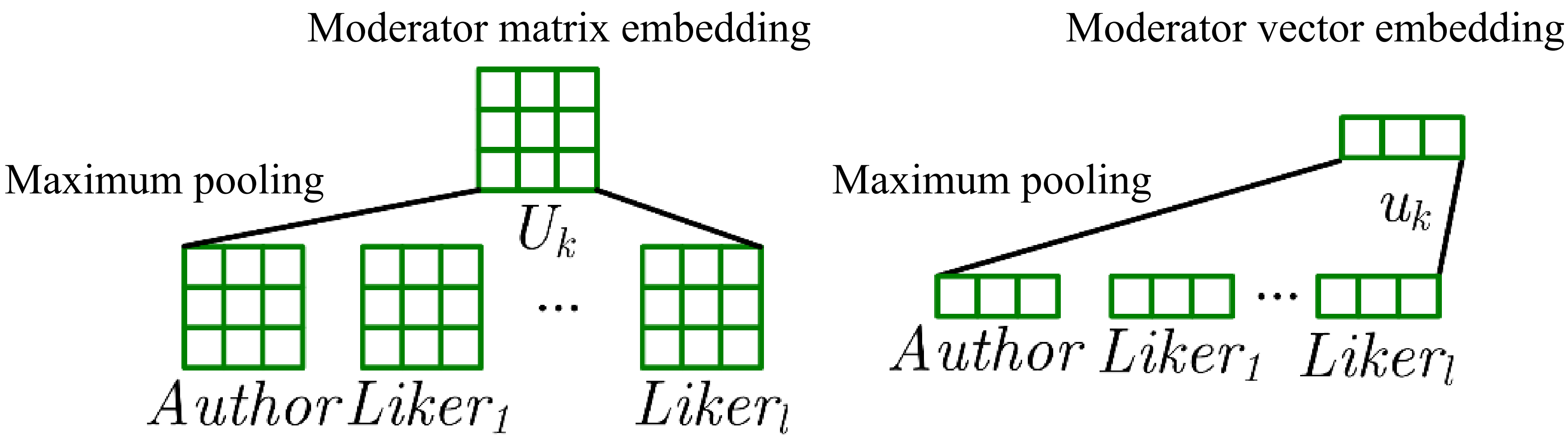
- U_k models the user reading preference for certain semantics
- T_j models the topic semantics



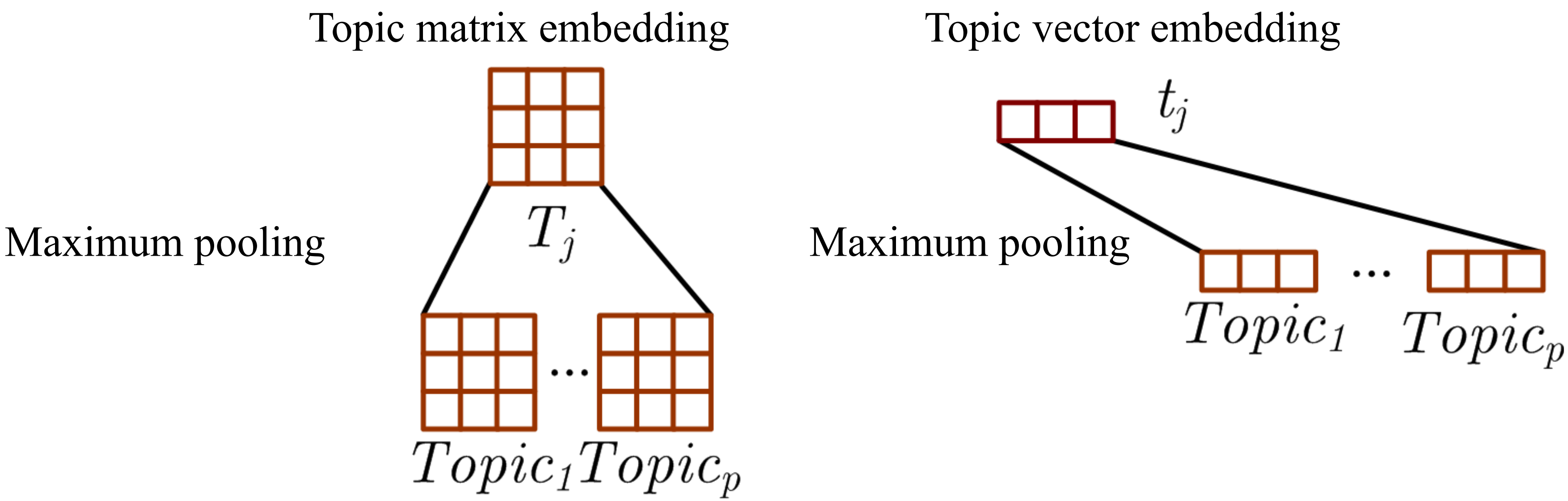
User- and topic-dependent stance tendency

- u_k models the user stance preference
- t_j models the topic stance tendency

Authors and Likers

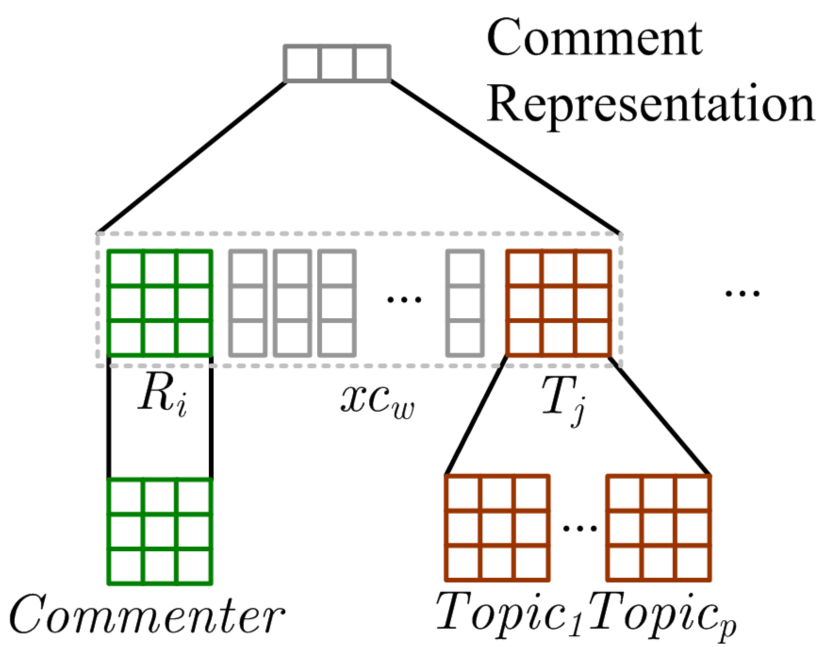


Topics

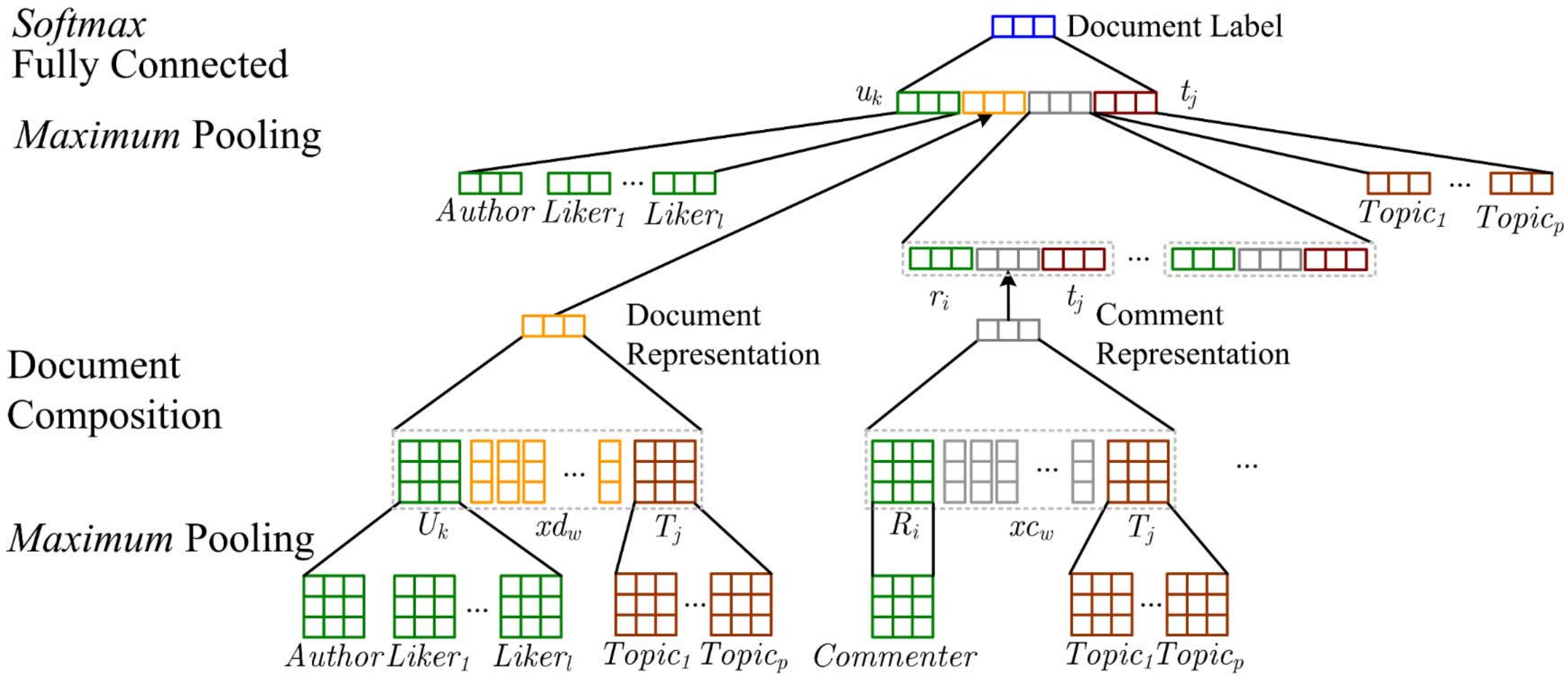

Topic matrix embedding
Maximum pooling
T_j
$Topic_1$ $Topic_p$
Topic vector embedding
t_j
Maximum pooling
$Topic_1$ $Topic_p$

Comment model

- Short document with only author



UTCNN – full view



Dataset

- Facebook fan groups
  - Author/liker/comment/commenter
  - Single topic (learn latent topics by LDA)
  - Unbalance
  - Chinese
- Create Debate
  - Author
  - Four topics
  - Balance
  - English

Dataset

Dataset	FBFans				CreateDebate							
Type					ABO		GAY		OBA		MAR	
	Sup	*Neu*	*Uns*	All	*F*	*A*	*F*	*A*	*F*	*A*	*F*	*A*
Training	7,097	19,412	245	26,754	770.4	622.4	700.8	400.0	420.8	367.2	355.2	145.6
Development	155	2,785	11	2,951	-	-	-	-	-	-	-	-
Testing	252	2,619	19	2,890	192.6	155.6	175.2	100.0	105.2	91.8	88.8	36.4
All	7,504	24,816	275	32,595	963.0	778.0	876.0	500.0	526.0	459.0	444.0	182.0

Annotation results of FBFans and CreateDebate dataset

Experiment settings

- Convolution filter window sizes: 1, 2, 3
- Word embedding dimension: 50
- User/topic matrix embedding size: 250 (5X50)
- User/topic vector embedding size: 10
- Latent topics: 100
- Maximum topics per document: 3

Results - FBFans

Method	Features				F-score			F_1^{SNU}
	Content	User	Topic	Comment	Sup	Neu	Uns	
Majority					.000	.841	.000	.280
SVM -UniBiTrigram	V			V	.610	.938	.156	.621
SVM -AvgWordVec	V			V	.526	.100	.165	.336
SVM -AvgWordVec (transformed)	V	V	V	V	.597	.963	.210	.642
CNN (Kim, 2014)	V			V	.726	.964	.222	.648
RCNN (Lai et al., 2015)	V			V	.628	.944	.096	.605
UTCNN – user	V		V	V	**.748**	**.973**	.000	.580
UTCNN – topic	V	V		V	.643	.944	.476	.706
UTCNN – comment	V	V	V		.632	.940	.480	.707
UTCNN shared user embedding	V	V	V	V	.625	.969	.531	.732
UTCNN (full)	V	V	V	V	.698	.957	**.571**	**.755***

Results - CreateDebate

Method	Features		Topics				AVG
	Text	User	ABO	GAY	OBA	MAR	
Majority			.549	.634	.539	.695	.604
SVM -UniBiTrigram	V		.592	.569	.565	.673	.600
SVM -AvgWordVec	V		.559	.637	.548	.708	.613
SVM -AvgWordVec (transformed)	V	V	.859	.830	.800	.741	.808
CNN (Kim, 2014)	V		.553	.636	.557	.709	.614
RCNN (Lai et al., 2015)	V		.553	.637	.534	.709	.608
ILP (Hasan and Ng, 2013a)	V		.614	.626	.581	.669	.623
ILP (Hasan and Ng, 2013a)	V	V	.749	.709	.727	.754	.735
CRF (Hasan and Ng, 2013b)	V	V	.747	.699	.711	.754	.728
PSL (Sridhar et al., 2015)	V	V	.668	.727	.635	.690	.680
UTCNN – topic	V	V	.824	**.851**	.743	**.814**	.808
UTCNN – user	V		.617	.627	.599	.685	.632
UTCNN (full)	V	V	**.878**	.850	**.857**	.782	**.842***

Conclusion

- We have proposed UTCNN incorporating user, topic, content and comment information for stance classification on social media texts.
- UTCNN learns user embeddings for all users with minimum active degree.
- Topic information obtained from the topic model or the pre-defined labels further improves the UTCNN model.
- Comment information provides additional clues for stance classification.
- We have shown that UTCNN achieves promising and balanced results.

Hand-on Session

Outline

- CSentiPackage: UTCNN
  - Introduction
  - Model
  - Results
- Hands on real data
  - Environment
  - Data preprocessing
  - Tools
    - NTUSD and ANTUSD
    - Cmorph and Chinese Opinion Treebank
    - CopeOpi
    - UTCNN

Environment

- Software
  - OS: Linux
  - Programming language
    - Java 6 or higher
    - python 2.7
      - Theano 0.8.2
      - Keras 1.0.3
      - sklearn
- Hardware
  - Graphic cards (deep learning)

Demo Environment

- CPU
  - Intel Xeon E5-2630 v3 ×2
- RAM
  - 64 GB
- OS
  - Ubuntu 14.04 LTS
- Graphic cards
  - Nvidia Tesla K40 ×2

Preprocessing

- Tokenize
  - Jieba
  - CKIP
  - Stanford parser
- Part-of-speech tagging
  - CKIP
  - Stanford parser

NTUSD

- National Taiwan University Sentiment Dictionary
- Release date: 2006
- Language: Traditional/ Simplified Chinese
- Data: 11,088 sentiment words
  - 2,812 positive words
  - 8,276 negative words

NTUSD – package

- 

NTUSD_negative_unicode.txt

刀刃
刁難
力盡
匕首
下地獄
下垂
下垂度
下流
下流的
下降
下陷
下等
下等的
下跌
下獄
下層社會

NTUSD_positive_unicode.txt

一帆風順的
一流
一致
一致的
了不起
了不起的
了解
人性
人性的
人格高尚
人格高尚的
人情
人情味
入神
入神的
入迷

NTUSD - reference

- Ku, L. W., Liang, Y. T., & Chen, H. H. (2006, March). Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *AAAI spring symposium: Computational approaches to analyzing weblogs*.
- http://doraemon.iis.sinica.edu.tw/coling2016_tutorial/downloads/NTUSD_traditional.zip
- http://doraemon.iis.sinica.edu.tw/coling2016_tutorial/downloads/NTUSD_simplified.zip

ANTUSD

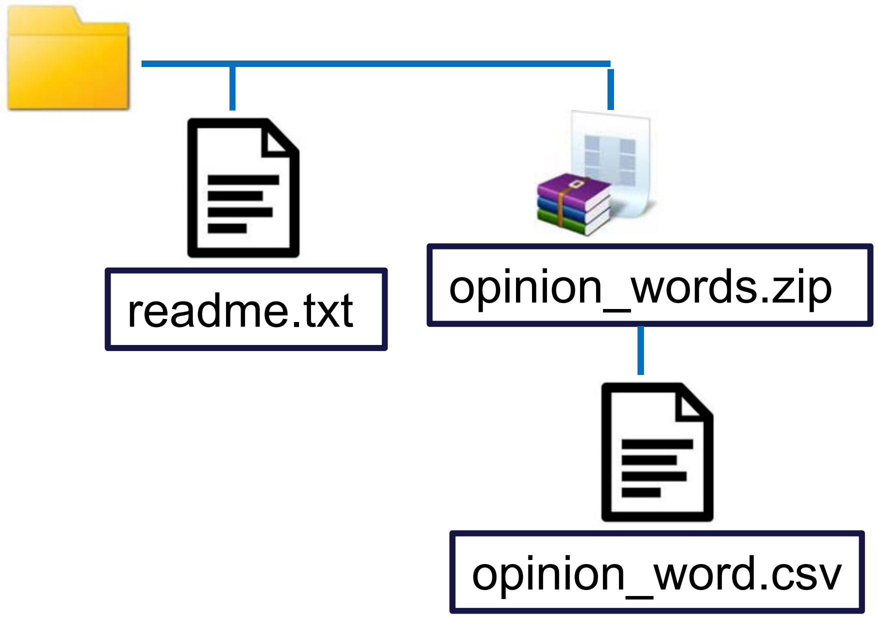
- Augmented NTUSD
- Release date: 2016
- Language: Traditional/ Simplified Chinese
- Data: 27,221 words
  - 9,382 positive words
  - 16 neutral words
  - 11,224 negative words
  - 5,415 non-opinion words
  - 612 negation words

ANTUSD - example

-

	Score	Pos	Neu	Neg	Nonop
支持 (support)	0.0381147	1	0	0	0
全力支持 (fully support)	0.2870457	1	0	0	0
不支持 (not support)	-0.1949018	0	0	1	0

ANTUSD - package

- 


```
一分打點,0.287740425,1,0,0,0,0
一反,-0.4540008,0,0,1,0,0
一反前態,-0.2520004,0,0,1,0,0
一夫當關,0.224588275,1,0,0,0,0
一巴掌,-0.221000867,0,0,1,0,0
一心想,0.293363867,1,0,0,0,0
一手造成,-0.174794575,0,0,1,0,0
一文不值,-0.1126813,0,0,1,0,0
一日千里,0.04946985,1,0,0,0,0
一片譁然,-0.00426785,0,0,1,0,0
一再,0.0,0,0,1,0,0
一再,0.0,1,0,0,0,0
一再出現,0.02988815,1,0,0,0,0
一再叮囑,-0.0085089,0,0,1,0,0
一再受挫,-0.1315382,0,0,1,0,0
一同競爭,0.049700275,1,0,0,0,0
一吐為快,0.0957688,1,0,0,0,0
一向,0.0540008,1,0,0,0,0
一向認為,0.063013275,1,0,0,0,0
一如,0.2082702,1,0,0,0,0
一如以往,0.080425,1,0,0,0,0
```

ANTUSD - reference

- Wang, Shih-Ming, and Lun-Wei Ku. "ANTUSD: A Large Chinese Sentiment Dictionary." in *LREC 2016*.
- http://doraemon.iis.sinica.edu.tw/coling2016_tutorial/downloads/ANTUSD_traditional.zip
- http://doraemon.iis.sinica.edu.tw/coling2016_tutorial/downloads/ANTUSD_unicode.zip

Cmorph

- Cmorph.txt: morphological types are labeled by numbers:
  - 1:Parallel
  - 2: Substantive-Modifier
  - 3: Subjective-Predicate
  - 4: Verb-Object
  - 5: Verb-Complement
  - 8: Others

刀械1
刁民2
力言2
力取2
力保2
力鬥2
力陳2
力圖2
三思2
三連2
三圍2
下去1
下台4
下市4
下田4
下挫2

*6: Negation and 7: Confirmation are detected by rules

*Huang, Ting-Hao, Ku, Lun-Wei and Chen, Hsin-Hsi, Predicting Morphological Types of Chinese Bi-Character Words by Machine Learning Approaches, *LREC* 2010, pages 844-850,

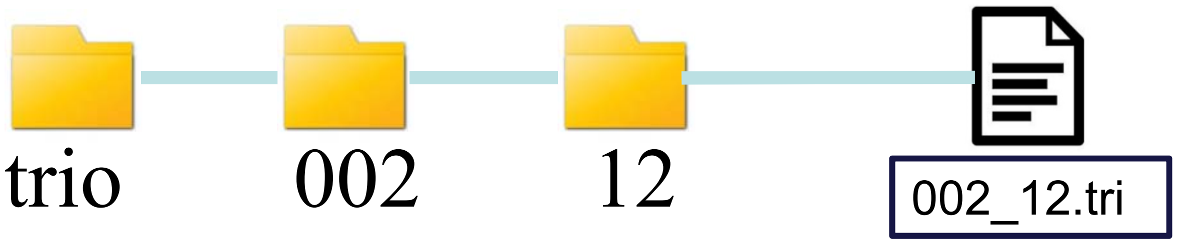
Chinese Opinion Treebank

- Excel file: sentence.csv

2	chtb_001.raw	2	N		
3	chtb_001.raw	3	Y	POS	ACTION
4	chtb_001.raw	4	Y	NEU	STATE
5	chtb_001.raw	5	Y	POS	STATE
6	chtb_001.raw	6	Y	POS	STATE
7	chtb_001.raw	7	Y	POS	STATE
8	chtb_001.raw	8	Y	POS	STATE
9	chtb_001.raw	9	N		
10	chtb_001.raw	10	Y	POS	STATE
11	chtb_001.raw	11	N		
12	chtb_002.raw	12	Y	POS	STATE
13	chtb_002.raw	13	N		
14	chtb_002.raw	14	N		
15	chtb_002.raw	15	N		
16	chtb_002.raw	16	Y	POS	STATE
17	chtb_002.raw	17	N		
18	chtb_002.raw	18	N		
19	chtb_002.raw	19	N		
20	chtb_002.raw	20	N		
21	chtb_003.raw	21	Y	POS	STATE
22	chtb_003.raw	22	N		
23	chtb_003.raw	23	Y	POS	STATE

*Ku, Lun-Wei, Huang, Ting-Hao and Chen, Hsin-Hsi, Construction of Chinese Opinion Treebank, *LREC* 2010, pages 1315-1319.

Chinese Opinion Treebank



(docID) (senID)

Chinese Opinion Treebank

- 外商 投資 企業
  成為 中國 外貿
  重要 增長點

<S ID=12>
((IP-HLN (NP-SBJ (NN外商)
(NN投資)
(NN 企業))
(VP (VV成為)
(NP-OBJ (NP (NP-PN (NR中國))
(NP (NN外貿)))
(ADJP (JJ重要))
(NP (NN增長點))))))
</S>

002_12.tree →

```
0:1,
1:2,6,
2:3,4,5,
3:
4:
5:
6:7,8,
7:
8:9,14,16,
9:10,12,
10:11,
11:
12:13,
13:
14:15,
15:
16:17,
17:
```

002_12.tri - 記事本

檔案(F) 編輯(E) 格式(O ... H)

```
0,8,15,17,2
1,6,7,8,4
2,1,2,6,3
```

CopeOpi - intro

- Unsupervised Chinese Sentiment scoring tool
- Dictionary: ANTUSD
- Language: Traditional Chinese
- Preprocessing
  - Tokenization
  - POS tagging (CKIP format)

CopeOpi – empirical usage

支持	核能	，	支持	核四
Support	nuclear power	,	support	Lungmen nuclear power plant
VC	Na	COMMA-CATEGORY	VC	Nc

，	享受	相對	便宜	的	電價	。
,	enjoy	relatively	cheaper		power rate	.
COMMA-CATEGORY	VJ	VH	VH	DE	Na	PERIOD-CATEGORY

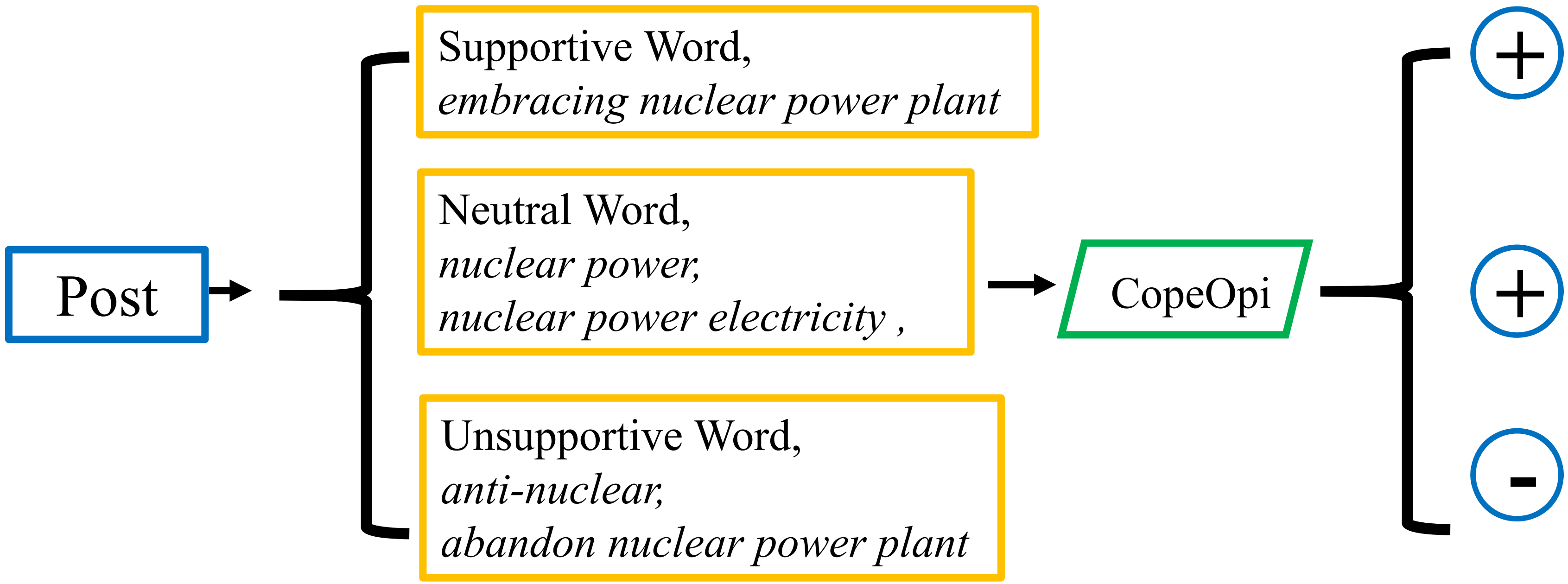
CopeOpi – empirical usage

支持	核能	，	支持	核四
Support	nuclear power	,	support	Lungmen nuclear power plant
0.0381147	0.0	0.0	0.0381147	0.0

，	享受	相對	便宜	的	電價	。
,	enjoy	relatively	cheaper		power rate	.
0.0	0.0340755	-0.042713	-0.3732	0.0	0.0	0.0

Document Score = 0.0675917

CopeOpi – transition process



Score = *Sup-Uns+Neu*

CopeOpi

- Package including
  - CopeOpi program, written in Java
  - CopeOpi source code
  - ANTUSD
  - A demo text
  - Read me

CopeOpi - package

- dic: dictionary files
- out: output folder
- CopeOpi.class (.java): interface
- OpinionCore_Enhanced.class (.java): core
- readme.txt: readme file
- file.lst: input file list
- test.txt: example input file
- run.sh: running script

CopeOpi – example

- $ ./run.sh
  - Run the CopeOpi with the files in the list “file.lst”

```
test.txt 0001
```

```
□  CopeOpi_EnhancedVersion  ./run.sh
Dictionaries Reload...
Processing: 0001
Analyzing Finish
```

- Check the results in out/0001.txt

```
支持/0.03811470000000001 核能/0.0 ，/0.0 支持/0.03811470000000001 核四/0 ，/0.0 享
64 便宜/-0.3732806 的/0.0 電價/0.0 。/0.0
支持核能，支持核四，享受相對便宜的電價。
***Score=0.06759174999999995
```

CopeOpi – example

- Result summary in ./out.csv

```
0001,0.06759174999999995,Positive
```

CopeOpi – reference

- CopeOpi
  - Ku, L. W., Ho, H. W., & Chen, H. H. (2009). Opinion mining and relationship discovery using CopeOpi opinion analysis system. Journal of the American Society for Information Science and Technology, 60(7), 1486-1503.
- CopeOpi with transition process
  - Chen, W. F., Ku, L. W., & Lee, Y. H. (2015). Mining Supportive and Unsupportive Evidence from Facebook Using Anti-Reconstruction of the Nuclear Power Plant as an Example. In 2015 AAAI Spring Symposium Series.
- http://doraemon.iis.sinica.edu.tw/coling2016_tutorial/downloads/CopeOpi_EnhancedVersion.zip

UTCNN - intro

- Aim
  - Stance Classification on Social Media
- Features
  - Information of social network platforms
    - Authorship
    - Likings
    - Topics
    - Comments

UTCNN - data

Field	Author and liker IDs	Topic IDs	Label	Content	Commenters	Comments
Delimiter	space	space			space	comma
Tokenize				space		space

UTCNN - data

- 3 46 57 … 573 49 61 4 -1 <sssss>福 島 核電廠 的 熔 毀 核 燃料棒 到底 有沒有 掉到 地下水層 …..<sssss>詳 見 俄國 時報 電視 專訪 <sssss> 544 490 565 … 428 危機 ,如果 安全 你 家 借放 ,事實 是 沒有 人 知道 真相 這 些 都 只是 推論 就 看 誰 的 推論 有 根據 合理 奇怪 的 是 擁核 五 毛 只 根據 東京 電力 的 說法 而 東京 電力 是 最 有 利益 關係 最 有 企圖 掩藏 事實 的 事主 貼 此 文 是 提 供 大家 獨立 沒有 核電 利益 纏身 的 核工 專家 與 小出裕 章 的 推論 僅 供 參考

UTCNN - package

- dataset: data required for this tutorial
  - data.train
  - data.dev
  - data.test
  - data.readme
  - vectors.50d.txt
- h5: parameters saved here
- pickle: results saved here
- config.ini: configuration file
- UTCNN_release.py: main program
- readme: readme file

UTCNN - example

- Package including
  - UTCNN model, written in python
  - Chinese word embeddings by GloVe
  - Demo data
    - 1000 training samples
    - 100 development samples
    - 100 testing samples

UTCNN - example

- $ python UTCNN_release.py config.ini

```
release  python UTCNN_release.py config.ini
Using Theano backend.
Using gpu device 0: Tesla K40c (CNMeM is enabled with initial size: 75.0% of memory, cuDNN 4007)
Load embedding file: ./dataset/vectors.50d.txt
Load Embedding Elapse:  1.00549221039
Load data file: ./dataset/data.train
Load data file: ./dataset/data.test
Load data file: ./dataset/data.dev
Load Data Elapse:  0.158798933029
Train Max:  (215, 2236, 4777)
Test Max:  (28, 1726, 2190)
Dev Max:  (38, 790, 2188)
max user: 7193
max comment: 100
max comment length: 628
user length: 37470
topic length: 99
Initialization Elapse:  1.22663116455
Sentences Processing Elapse:  3.40131402016
Train on 1000 samples, validate on 72 samples

 660/1000 [==================>...........] - ETA: 31s - loss: 0.5064 - acc: 0.7788
```

UTCNN - example

```
Epoch 1/10
1000/1000 [==============================] - 98s - loss: 0.4589 - acc: 0.8020 - val_loss: 0.5727 - val_acc: 0.7222
Epoch 2/10
1000/1000 [==============================] - 98s - loss: 0.1708 - acc: 0.9530 - val_loss: 0.6982 - val_acc: 0.7222
Epoch 3/10
1000/1000 [==============================] - 98s - loss: 0.0789 - acc: 0.9820 - val_loss: 0.7685 - val_acc: 0.7083
Epoch 4/10
1000/1000 [==============================] - 98s - loss: 0.0457 - acc: 0.9880 - val_loss: 0.8213 - val_acc: 0.6667
Epoch 5/10
1000/1000 [==============================] - 98s - loss: 0.0316 - acc: 0.9910 - val_loss: 0.8169 - val_acc: 0.6667
Epoch 6/10
1000/1000 [==============================] - 98s - loss: 0.0234 - acc: 0.9960 - val_loss: 0.8530 - val_acc: 0.6806
Fitting Elapse:  1436.94893384
/usr/local/lib/python2.7/dist-packages/sklearn/metrics/classification.py:1074: UndefinedMetricWarning: Precision and F-score are ill-defined
 and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python2.7/dist-packages/sklearn/metrics/classification.py:1076: UndefinedMetricWarning: Recall and F-score are ill-defined an
d being set to 0.0 in labels with no true samples.
  'recall', 'true', average, warn_for)
(array([ 0.78571429,  0.95774648,  0.        ]), array([ 0.78571429,  0.95774648,  0.        ]), array([ 0.78571429,  0.95774648,  0.
 ]), array([14, 71,  0]))
```

- Parameters: ./h5/
  - Best: UTCNN_best.h5
  - Others: UTCNN_itr[00].h5
- Prediction results: ./pickle/predict.pickle

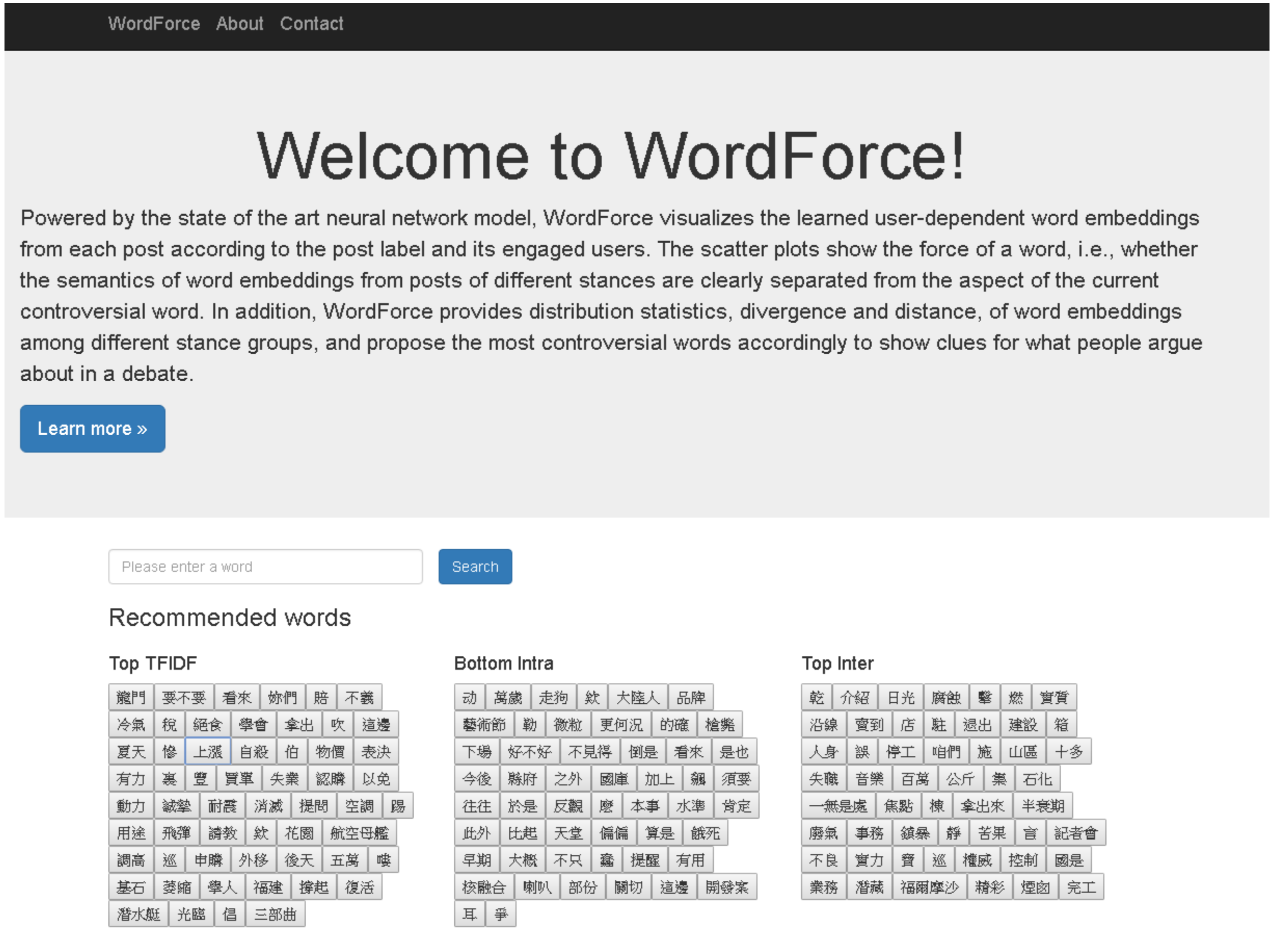
UTCNN - example

- config.ini

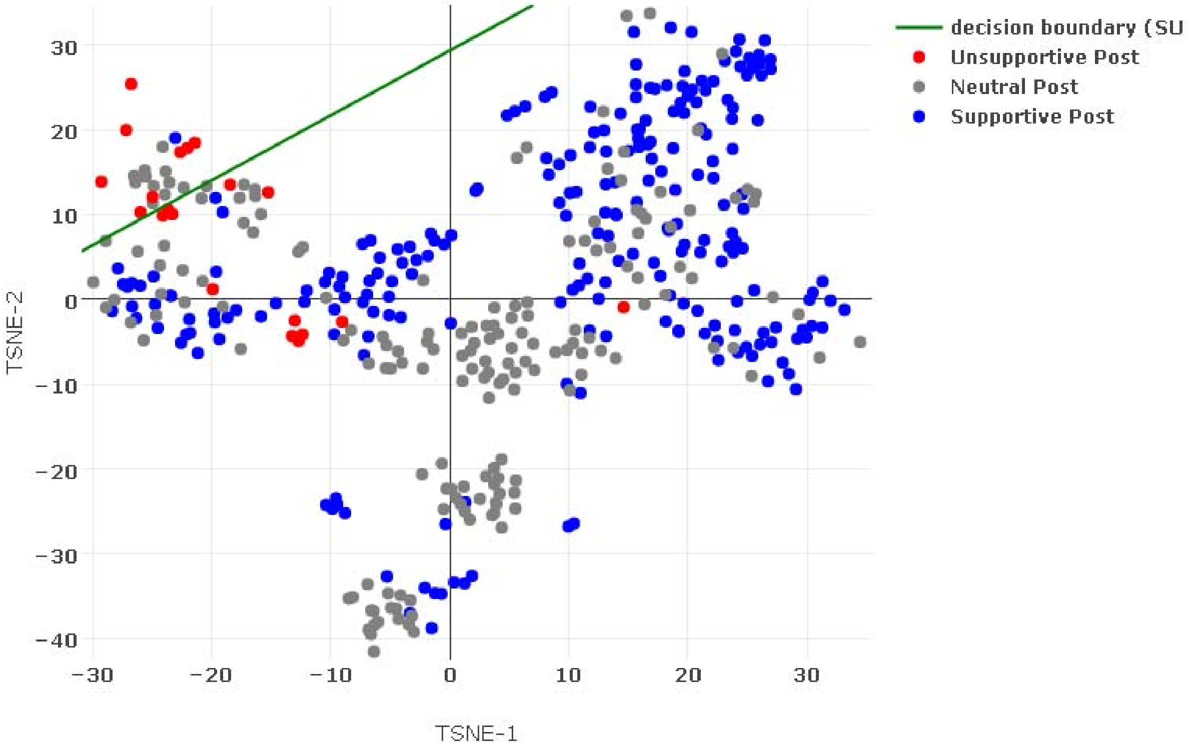
```
[Files]
embedding_file = ./dataset/vectors.50d.txt
# embedding files, trained by GloVe
train_file = ./dataset/data.train
# input file for training, one sample per line
test_file = ./dataset/data.test
# input file for testing, one sample per line
dev_file = ./dataset/data.dev
# input file for development, one sample per line
save_each = ./h5/UTCNN_itr{epoch:02d}.h5
# saved filename of each iteration
save_final = ./h5/UTCNN_best.h5
# saved fileanme for the final iteration
save_pickle = ./pickle/predict.pickle
# saved fileanme for the prediiction, saved in pickle format

[Pars]
v_dim = 50
# dimension in the word embedding file
u_dim = 10
# dimension of the user vector embeddings
mini_u_dim = 5
# first dimension of the user matrix embeddings
t_dim = 10
# dimension of the topic vector embeddings
```

UTCNN - demo


WordForce About Contact
Welcome to WordForce!
Powered by the state of the art neural network model, WordForce visualizes the learned user-dependent word embeddings from each post according to the post label and its engaged users. The scatter plots show the force of a word, i.e., whether the semantics of word embeddings from posts of different stances are clearly separated from the aspect of the current controversial word. In addition, WordForce provides distribution statistics, divergence and distance, of word embeddings among different stance groups, and propose the most controversial words accordingly to show clues for what people argue about in a debate.
Learn more »
Please enter a word
Search
Recommended words
Top TFIDF
龍門 要不要 看來 妳們 賠 不義
冷氣 稅 絕食 學會 拿出 吹 這邊
夏天 慘 上漲 自殺 伯 物價 表決
有力 裹 豐 買單 失業 認購 以免
動力 誠摯 耐震 消滅 提問 空調 踢
用途 飛彈 請教 欸 花園 航空母艦
調高 巡 申購 外移 後天 五萬 嘍
基石 萎縮 學人 福建 撐起 復活
潛水艇 光臨 信 三部曲
Bottom Intra
动 萬歲 走狗 欸 大陸人 品牌
藝術節 勒 微粒 更何況 的確 槍斃
下場 好不好 不見得 倒是 看來 是也
今後 縣府 之外 國庫 加上 飆 須要
往往 於是 反觀 麼 本事 水準 肯定
此外 比起 天堂 偏偏 算是 餓死
早期 大概 不只 蠶 提醒 有用
核融合 喇叭 部份 關切 這邊 開發案
耳 爭
Top Inter
乾 介紹 日光 腐蝕 擊 燃 實質
沿線 賣到 店 駐 退出 建設 箱
人身 誤 停工 咱們 施 山區 十多
失職 音樂 百萬 公斤 黨 石化
一無是處 焦點 棟 拿出來 半衰期
廢氣 事務 鎮暴 靜 苦果 言 記者會
不良 實力 寶 巡 權威 控制 國是
業務 潛藏 福爾摩沙 精彩 煙囪 完工

UTCNN - demo

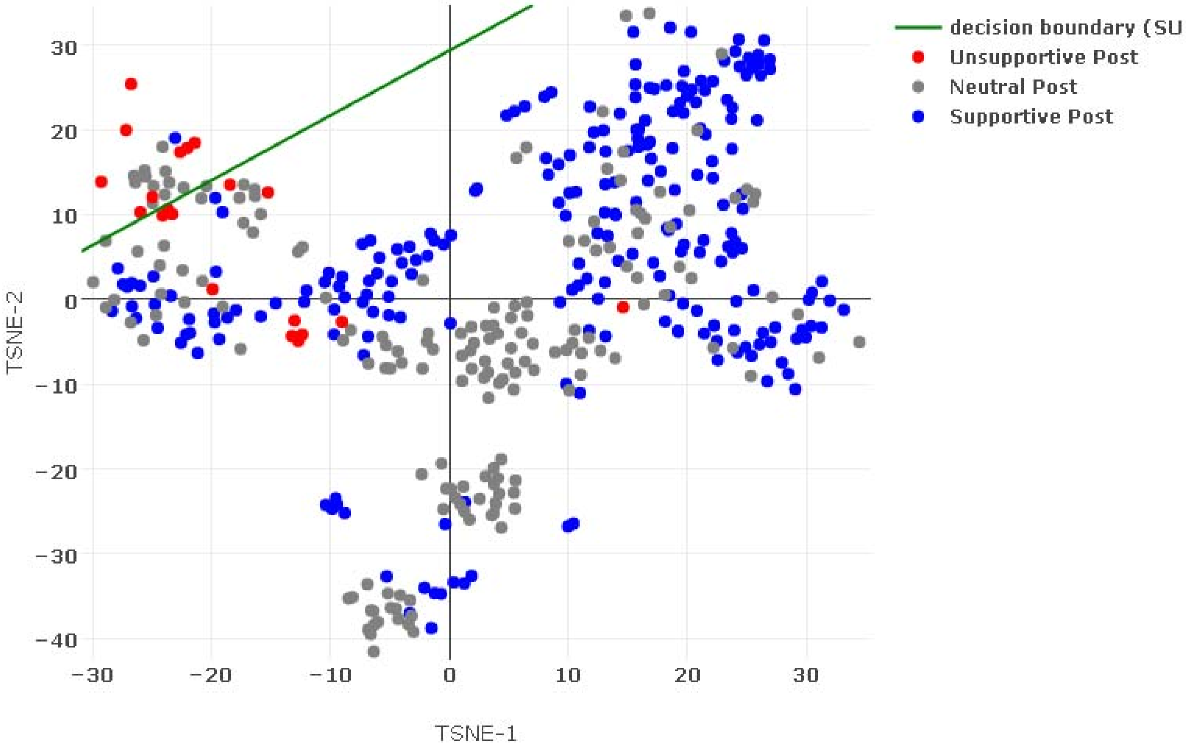


Distribution Statistics

Intra-group	Distance
Supportive	1.20
Neutral	1.51
Unsupportive	0.99
Inter-group	**Distance**
Supportive/Unsupporitve	2.34
Neutral/Supportive	1.87
Neutral/Unsupportive	1.43

http://doraemon.iis.sinica.edu.tw/wordforce/

UTCNN - demo



Distribution Statistics

Intra-group	Distance
Supportive	1.20
Neutral	1.51
Unsupportive	0.99
Inter-group	**Distance**
Supportive/Unsupporitve	2.34
Neutral/Supportive	1.87
Neutral/Unsupportive	1.43

世上哪種發電沒有汙染或噪音?不是噪音就汙染不然就是發電量不夠@@說句難聽的~你們這反核部隊裡面真的反核的有幾個?裡面不知加了多少政治狂?要不是接近選舉了這些人會加入你們嗎?真要反核嗎?好!!要反核第一條件~~廢核第一要面對的違約金一千多億由這些反核廢核人士負責!!!這樣我就同意反核@@不要全民買單由你們這些反核負責!!!敢不敢???關於林XX不是要絕食嗎?幹嘛還去醫院?老招式了@@換招新招式吧!看了就煩!看也知道選舉到了.......老梗@@再來就是幹嘛罵警察!莫非你們和林X帆陳X廷一樣?需要警察叫警察保護!!!不要警察就說警察暴力!打人!沒人性!!!!你們解散不就沒這事情了!一個人被你們浪費一小時全台北市多少人???多少個一小時???不是不可遊行抗議~但是你們不要妨礙他人啊!!!將心比心啊!!!!!有本事就不要刪除我貼文讓大家看!!!!!

UTCNN - reference

- Wei-Fan Chen and Lun-Wei Ku. (2016). UTCNN: a Deep Learning Model of Stance Classification on Social Media Text. In COLING 2016 main track.
- Wei-Fan Chen, Fang-Yu Lin and Lun-Wei Ku. (2016). WordForce: Visualizing Controversial Words in Debates. In COLING 2016 demo track.
- http://doraemon.iis.sinica.edu.tw/coling2016_tutorial/downloads/UTCNN_release_161114.zip

Conclusion

- Chinese sentiment dictionaries
- Lexicon-based and deep learning-based models for sentiment analysis
- The utilization of these resources and tools

Final Wrap Up

- Basic concepts of sentiment analysis and Chinese text processing
- Introduction of CSentiPackage
- Hand-on CSentiPackage

Now you should be able to work with your Chinese texts and detect sentiment from them!

Something Important About CSentiPackage

- CSentiPackage you obtained here is only for your group to use for the research purpose.
- Part of it has been officially released so they can be downloaded any time.
- To obtain the other, **join the next CSentiPackage tutorial** or check what's new @ http://academiasinicanlplab.github.io/

Join Our Three Demos Here

December 15th,10:30–12:30 Demo Session 3

1. Sensing Emotions in Text Messages: An Application and Deployment Study of EmotionPush

December 16th,14:00–15:30 Demo Session 6

2. WordForce: Visualizing Controversial Words in Debates

3. Automatically Suggesting Example Sentences of Near-Synonyms for Language Learners

from

Lun-Wei Ku & Wei-Fan Chen

NLPSA Lab, Academia Sinica