Mining Informal Language from Chinese Microtext : Joint Word Recognition and Segmentation

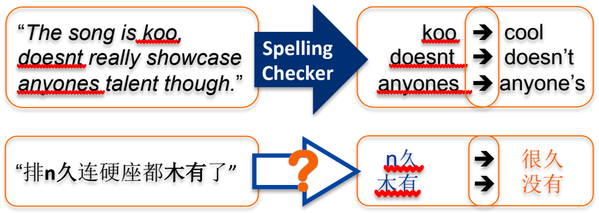
Aobo Wang and Min-Yen Kan

NUS National University of Singapore

Web IR / NLP Group
Interactive and Digital Media Institute
{wangaobo,kanmy}@comp.nus.edu.sg

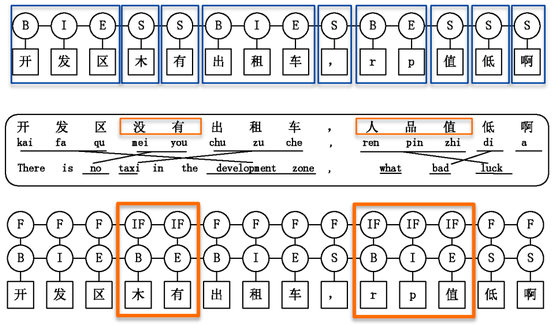
Introduction

- We propose to jointly model the two tasks of Informal word recognition (**IWR**) and Chinese word segmentation (**CWS**)
- Informal words in Chinese are difficult to recognize (shown in Figure 1) because they:
  - Are not indicated by word delimiters
  - Consist of a mix of numbers, alphabetic letters and Chinese characters



While tools like spell checking may work to link informal English words to their formal counterpart, they don't work for Chinese microtext ("tweet" / Weibo)

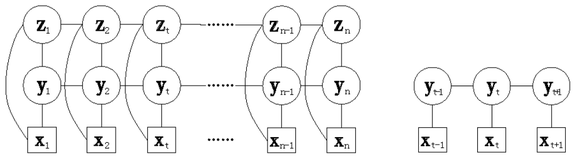
Problem Formalization



A Chinese microtext (in squares) with labels (in circles).
F/IF indicates the character as part of a **formal/informal** word
BIES is the widely-used coding scheme for segmentation

- Incorrect segmentation (in blue rectangles) caused by informal words (in orange rectangles)
- Segmentations to neighbors help recognize informal words
- CWS and IWR are mutually dependent
- Formulate as a 2-layer sequential labelling task

2-Layer Factorial CRF Model



Graphical representations of the two types of CRFs used in this work. $\mathbf{y}_t$ denotes the 1st layer label, $\mathbf{z}_t$ denotes the 2nd layer label, and $\mathbf{x}_t$ denotes the observation sequence.

- FCRF:
  - Introduces a pairwise factor among different variables at each position
  - captures the joint distribution among layers
- Compared with LCRF:
  - FCRF has fewer parameters
  - FCRF needs less training data

Experiment Results

FCRF versus baselines on CWS. '†' ('*') indicates statistical significance at p<0.001 (0.05) compared with the previous row.

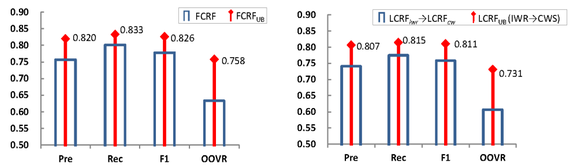
CWS	Pre	Rec	F_1	OOVR
HHMM (ICTCLAS, 2011)	0.640	0.767	0.698	0.551
LCRF (Sun and Xu, 2011)	0.661‡	0.691‡	0.675	0.572‡
$LCRF_{iwr} \rightarrow LCRF_{cws}$	0.741‡	0.775‡	0.758*	0.607*
FCRF	**0.757‡**	**0.801‡**	**0.778***	**0.633***

- Microtext is difficult to segment
- CWS benefits significantly from the results of IWR
- Joint inference works best

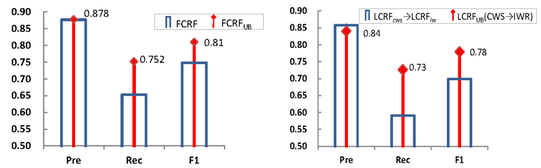
FCRF versus baselines on IWR. '†' ('*') indicates statistical significance at p<0.001 (0.05) compared with the previous row.

IWR	Pre	Rec	F_1
SVM (Xia and Wong, 2008)	0.382	0.621	0.473
DT	0.402*	0.714*	0.514*
$LCRF_{cws} \rightarrow LCRF_{iwr}$	0.858‡	0.591‡	0.699‡
FCRF	**0.877***	**0.655***	**0.750***

- SVM and DT tend to over predict informality
- IWR task is improved significantly with CWS tasks
- Joint inference again is most effective



Upper bound systems versus their counterparts on CWS.



Upper bound systems versus their counterparts on IWR.

- Still room for improving CWS with better IWR
- FCRF makes significant progress towards the UB
- Again, can further improve IWR with better CWS
- CWS enables IWR to make more predictions

Feature set evaluation. FCRF-new refers to the system without the novel features we introduced, that are marked with "*".

	CWS (F_1)	IWR(F_1)
FCRF-new	0.690	0.552
FCRF	**0.778***	**0.748***

- Lexical Features
- Dictionary-based Features*
- Statistical Features*

FCRF versus Adapted SVM for Joint Classification (SVM-JC). SVM-JC classifies input into the space of cross-product of the 2-layer labels.

	CWS (F_1)	IWR(F_1)
SVM	—	0.473
SVM-JC	0.711	0.624‡
FCRF	**0.778***	**0.748***

- Over-prediction is lessened
- FCRF is still more effective

Error Analysis

- Partially-observed informal words
  - "狠" ("很", "very") is a known informal word
  - "狠久" ("很久","for a long time") is informal
- Extremely short sentences
  - "肥家！太累了。。。"
  ("回家！太累了。。。","Go home! Exhausted.")
  - The informal word itself forms a short sentence
  - Two sentences are pragmatically related
  - But lexical dependency is weak
- Freestyle Chinese Named Entities

Freestyle Named Entity	Explanation
"榴莲雪媚娘"	"榴莲" ("durian"), "雪" ("snow"), "媚娘" ("charming lady")
"棉宝"	short for the cartoon name "海绵宝宝"
"dj文祥" "徐pp"	Usernames mixed of Chinese and alphabetic characters

Conclusion

- We evaluate our method on a manually-constructed data set with crowdsourced annotation
- The FCRF model yields significantly better performance than individual or sequential solutions
- We introduced novel features that Improve the performance significantly
- Upper bound systems validate the necessity and effectiveness of modeling the two tasks jointly