Weak Semi-Markov CRFs for NP Chunking in Informal Text

Aldrian Obaja Muis and Wei Lu

Singapore University of Technology and Design

Paper Contributions

In this paper, we contributed:

1. Noun phrase-annotated SMS corpus[1]

[1] Tao Chen and Min-Yen Kan (2013). "Creating a live, public short message service corpus: the NUS SMS corpus". In: *Language Resources and Evaluation*. Vol. 47. Springer Netherlands, pp. 299–335.

Paper Contributions

In this paper, we contributed:

1. Noun phrase-annotated SMS corpus[1]
2. Weak semi-Markov CRF

[1] Tao Chen and Min-Yen Kan (2013). "Creating a live, public short message service corpus: the NUS SMS corpus". In: *Language Resources and Evaluation*. Vol. 47. Springer Netherlands, pp. 299–335.

NP-annotated SMS Corpus

NP-annotated SMS Corpus

We used Brat Rapid Annotation Tool (BRAT)[2] for annotations, recruiting undergraduate students to annotate the noun phrases.

[2] http://brat.nlplab.org/

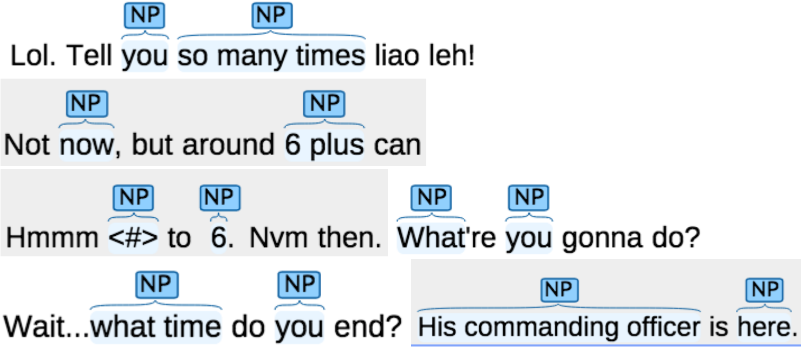
NP-annotated SMS Corpus

We used Brat Rapid Annotation Tool (BRAT)[2] for annotations, recruiting undergraduate students to annotate the noun phrases. Examples:

[2] http://brat.nlplab.org/

NP-annotated SMS Corpus

We used Brat Rapid Annotation Tool (BRAT)[2] for annotations, recruiting undergraduate students to annotate the noun phrases. Examples:



Lol. Tell you so many times liao leh!

Not now, but around 6 plus can

Hmmm <#> to 6. Nvm then. What're you gonna do?

Wait...what time do you end? His commanding officer is here.

[2] http://brat.nlplab.org/

Annotations Statistics

64 annotators

Annotations Statistics

64	annotators
26,500	SMS messages

Annotations Statistics

64	annotators
26,500	SMS messages
76,490	noun phrases

Annotations Statistics

64	annotators
26,500	SMS messages
76,490	noun phrases
359,009	tokens

Models

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

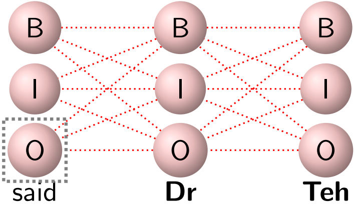


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

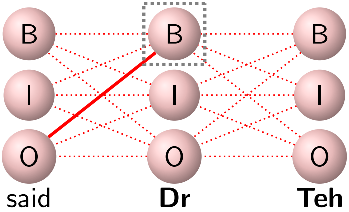


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

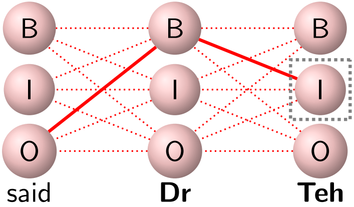


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

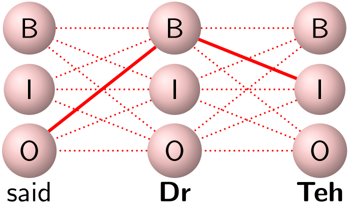


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

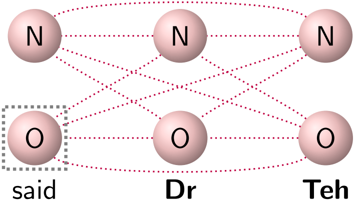


Fig. 2: Semi-CRF: $O(nL\,|\mathcal{Y}|^2)$

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

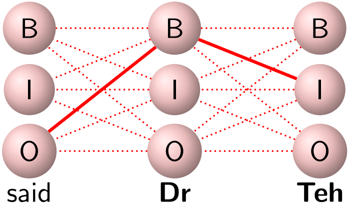


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

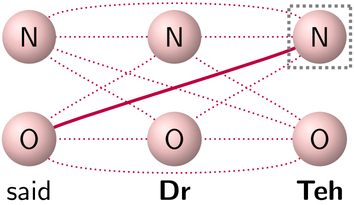


Fig. 2: Semi-CRF: $O(nL\,|\mathcal{Y}|^2)$

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

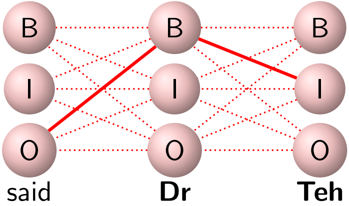


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

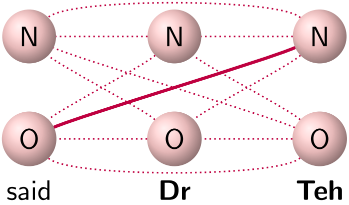


Fig. 2: Semi-CRF: $O(nL\,|\mathcal{Y}|^2)$

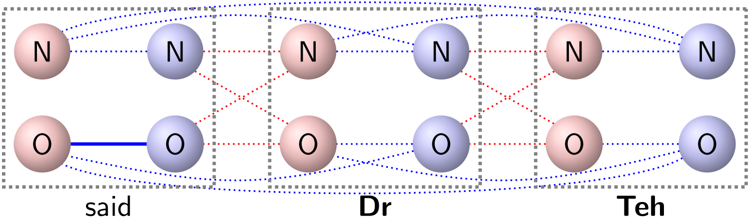


Fig. 3: Weak Semi-CRF: $O(n\,|\mathcal{Y}|^2 + nL\,|\mathcal{Y}|)$

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

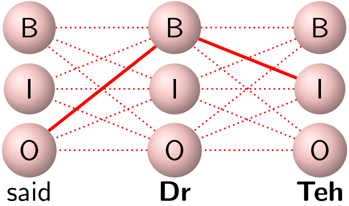


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

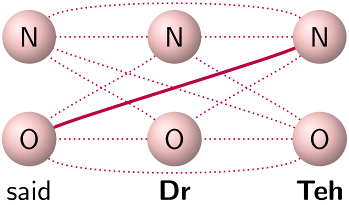


Fig. 2: Semi-CRF: $O(nL\,|\mathcal{Y}|^2)$

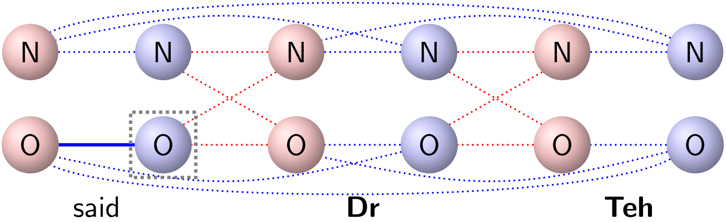


Fig. 3: Weak Semi-CRF: $O(n\,|\mathcal{Y}|^2 + nL\,|\mathcal{Y}|)$

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

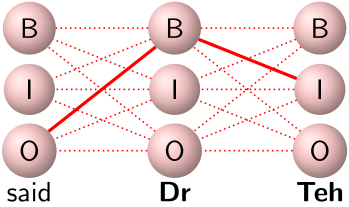


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

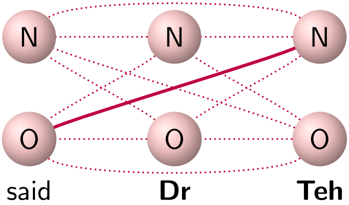


Fig. 2: Semi-CRF: $O(nL\,|\mathcal{Y}|^2)$

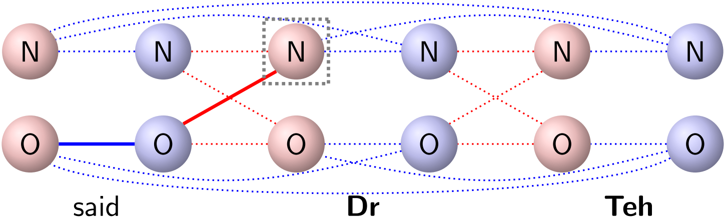


Fig. 3: Weak Semi-CRF: $O(n\,|\mathcal{Y}|^2 + nL\,|\mathcal{Y}|)$

Models Comparison

n : # words in the sentence, $|\mathcal{Y}|$: # labels, L : max segment length

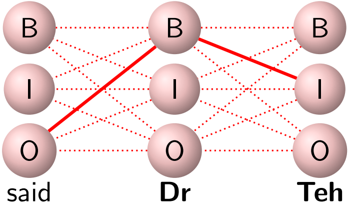


Fig. 1: Linear CRF: $O(n|\mathcal{Y}|^2)$

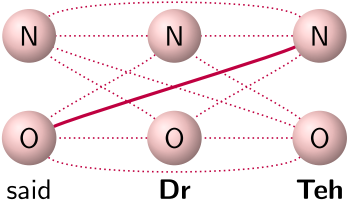


Fig. 2: Semi-CRF: $O(nL\,|\mathcal{Y}|^2)$

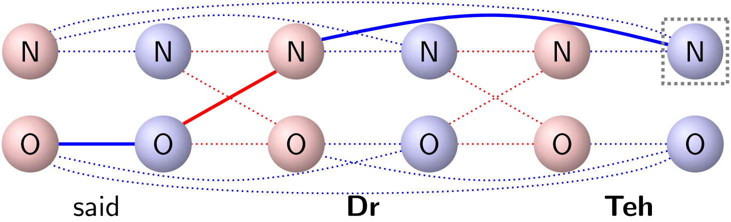
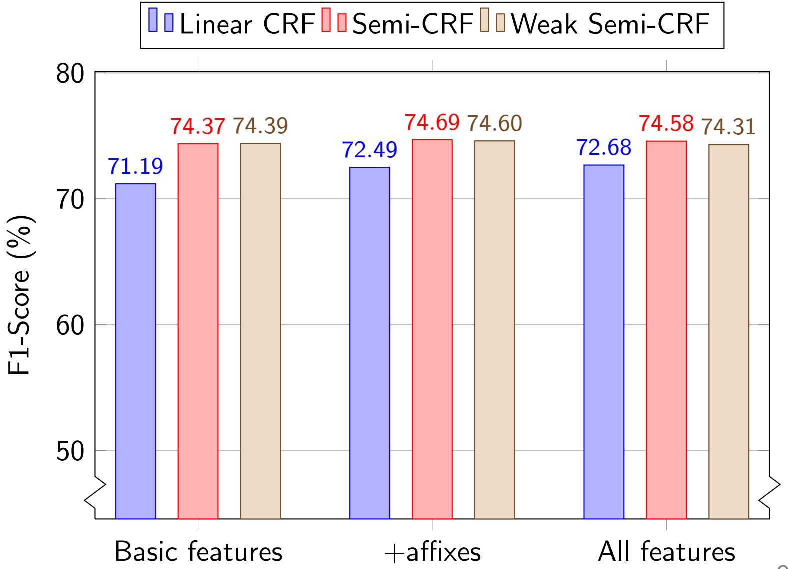


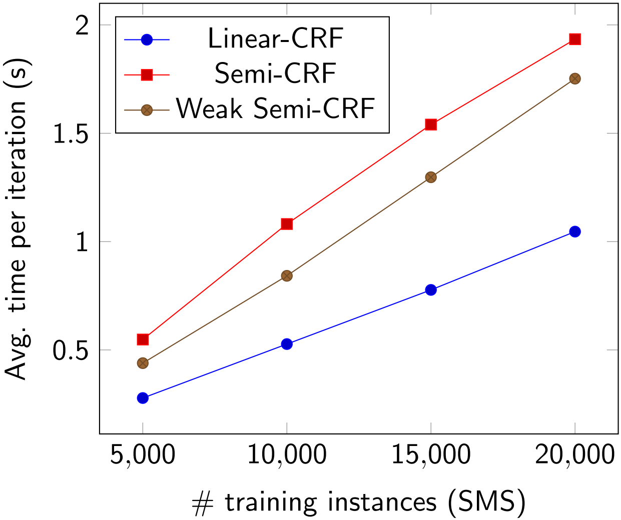
Fig. 3: Weak Semi-CRF: $O(n\,|\mathcal{Y}|^2 + nL\,|\mathcal{Y}|)$

Empirical Verification

F1-Score


Linear CRF
Semi-CRF
Weak Semi-CRF
F1-Score (%)
80
70
60
50
71.19
74.37
74.39
72.49
74.69
74.60
72.68
74.58
74.31
Basic features
+affixes
All features

Training Speed



Conclusion

Conclusion

- We have created a new NP-annotated dataset on informal text

Conclusion

- We have created a new NP-annotated dataset on informal text
- We can split the decisions of selecting segment length and segment type to improve the training time, while maintaining similar accuracy

Thank You

Code and data available at:
`http://statnlp.org/research/ie/`

Aldrian Obaja Muis and **Wei Lu**
Singapore University of Technology and Design