Filtered Semi-Markov CRF

Urchade Zaratiana, Nadi Tomeh, Niama El Khbir,
Pierre Holat, Thierry Charnois
1 LIPN, CNRS UMR 7030, France, 2 FI Group
{zaratiana,tomeh,elkhbir,holat,charnois}@lipn.fr
EMNLP 2023 (Findings)

1) Input text:

Alain Farley works at McGill University
PER ORG

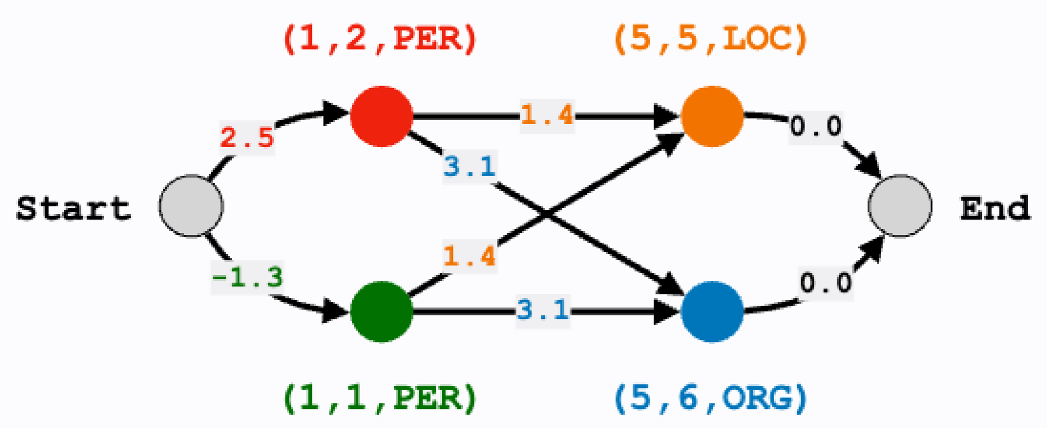
2) Filtering using ϕ_{local}:

(1,1,PER) (5,6,ORG)
(1,2,PER) (5,5,LOC)

3) Score segments using ϕ_{global}:

(1,1,PER) -1.3 (5,6,ORG) 3.1
(1,2,PER) 2.5 (5,5,LOC) 1.4

4) Graph construction:



5) Loss/decoding:

All possible paths and their scores:
Start (1,2,PER) (5,5,LOC) End -> 3.9
Start (1,2,PER) (5,6,LOC) End -> 5.6
Start (1,1,PER) (5,5,LOC) End -> 0.1
Start (1,1,PER) (5,6,LOC) End -> 1.8

Partition function Z(x):
Z=exp(3.9)+exp(5.6)+exp(0.1)+exp(1.8)

Decoding:
Return the path with maximum score

Introduction

	PER				ORG			LOC
Input sentence:	Alain	Farley	works	at	McGill	University	in	Montreal
BIO Tags:	B-PER	I-PER	O	O	B-ORG	I-ORG	O	B-LOC

Spans: (0,1,PER) (4,5,ORG) (7,7,LOC) (0,0,O) (1,1,O) (1,3,O) (3,3,O) ...

Entity Spans — Non-entity spans (null)

Background

1) Probabilistic Structured Prediction

Probability of a structure y given an input x.

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \boldsymbol{S}_\theta(\boldsymbol{y}|\boldsymbol{x})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{x})} \exp \boldsymbol{S}_\theta(\boldsymbol{y}'|\boldsymbol{x})}$$

Loss function (NLL)

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) &= -\log p_\theta(\boldsymbol{y}|\boldsymbol{x}) \\ &= -\boldsymbol{S}_\theta(\boldsymbol{y}|\boldsymbol{x}) + \log \mathcal{Z}_\theta(\boldsymbol{x})\end{aligned}$$

Inference: produce the most likely structure

$$\boldsymbol{y}^* = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})}{\arg\max}\, \boldsymbol{S}_\theta(\boldsymbol{y}|\boldsymbol{x})$$

2) Conditional Random Field (CRF) [1]

$$S(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{|\boldsymbol{x}|} \psi(y_i|\boldsymbol{x}) + \sum_{i=2}^{|\boldsymbol{x}|} \boldsymbol{T}[y_{i-1}, y_i]$$

- Token-level modelling + 1st order markov transition between tags
- Linear complexity
- Limited to modeling relationships between individual tokens

3) Semi-Markov CRF [2]

$$S(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{M} \phi(s_k|\boldsymbol{x}) + \boldsymbol{T}[l_{k-1}, l_k]$$

- (Span) segment-level modelling
  - **y** completely should cover the input sequence without overlapping
  - non-entity segments ('O' or null segments) have unit length
    - x = Alain Farley works at McGill University
    - y = (1,2,PER) (3,3,0) (4,4,0) (5,6,ORG)

- Higher-level segment features
- Slow training and inference
- Multiple valid paths (redundancy problem)
- Poor performance in practice

Filtered Semi-CRF

Main idea:

- Filter spans/segments using a lightweight classifier (easily parallelizable)
- Run Semi-CRF on the filtered spans
- Filtering is learned jointly with the Semi-CRF during training (multitask learning)
- No redundancy problem / Linear complexity (decoding)
- Strong performance (a variant our model won Arabic NER shared task)

Results

- We conducted experiments on flat NER datasets using BERT for token representation.

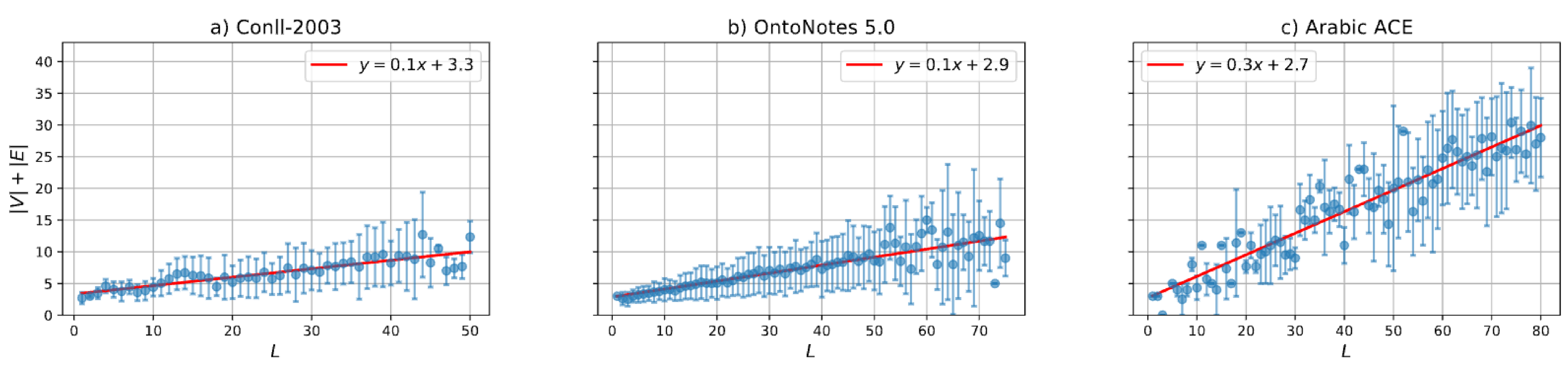
Models	CoNLL-2003			OntoNotes 5.0			Arabic ACE		
	P	R	F	P	R	F	P	R	F
Yu et al. (2020)	93.7	93.3	93.5	91.1	91.5	91.3	-	-	-
Yan et al. (2021)	92.61	93.87	93.24	89.99	90.77	90.38	-	-	-
Zhu and Li (2022)	93.61	93.68	93.65	91.75	91.74	**91.74**	-	-	-
Shen et al. (2022)	93.29	92.46	92.87	91.43	90.73	90.96	-	-	-
Zaratiana et al. (2022a)	94.29	93.33	93.81	90.21	91.21	90.71	85.35	83.64	**84.49**
El Khbir et al. (2022)	-	-	-	-	-	-	84.42	84.05	84.23
Our experiments									
CRF	93.29	92.21	92.75	89.00	90.16	89.57	82.79	84.44	83.61
Semi-CRF	92.37	90.49	91.42	88.91	89.78	89.34	82.97	84.24	83.60
+ Unit size null†	92.08	91.41	91.74	89.17	89.76	89.47	83.35	83.62	83.48
FSemiCRF	94.72	93.09	**93.89**	90.69	91.31	91.00	83.43	85.51	**84.46**
– w/o $\mathcal{L}_{global}$ (14)†	94.24	92.70	93.46	90.85	89.57	90.21	83.73	83.56	83.64

Inference speed

- **Walll clock time**

	CoNLL-2003 (|Y| = 4)			*OntoNotes 5.0 (|Y| = 18)*			*Arabic ACE (|Y| = 7)*		
	CRF	Semi-CRF	FSemiCRF	CRF	Semi-CRF	FSemiCRF	CRF	Semi-CRF	FSemiCRF
Scoring	3.9	3.9	3.9	4.8	4.9	4.9	8.1	8.3	8.3
Decoding	2.7	3.7	0.2	4.4	27.5	0.2	6.0	10.1	0.3
Decoding Speedup	1.3x	1.0x	**18.5x**	6.2x	1.0x	**137x**	1.7x	1.0x	**33.7x**
Overall	6.6	7.6	4.1	9.2	32.4	5.1	14.1	18.4	8.6
Overall Speedup	1.1x	1.0x	**1.8x**	3.5x	1.0x	**6.3x**	1.30x	1.0x	**2.1x**

- **Graph size after training**



[1] Lafferty et al. (2003) **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**

[2] Sarawagi & Cohen (2004) **Semi-Markov Conditional Random Fields for Information Extraction**