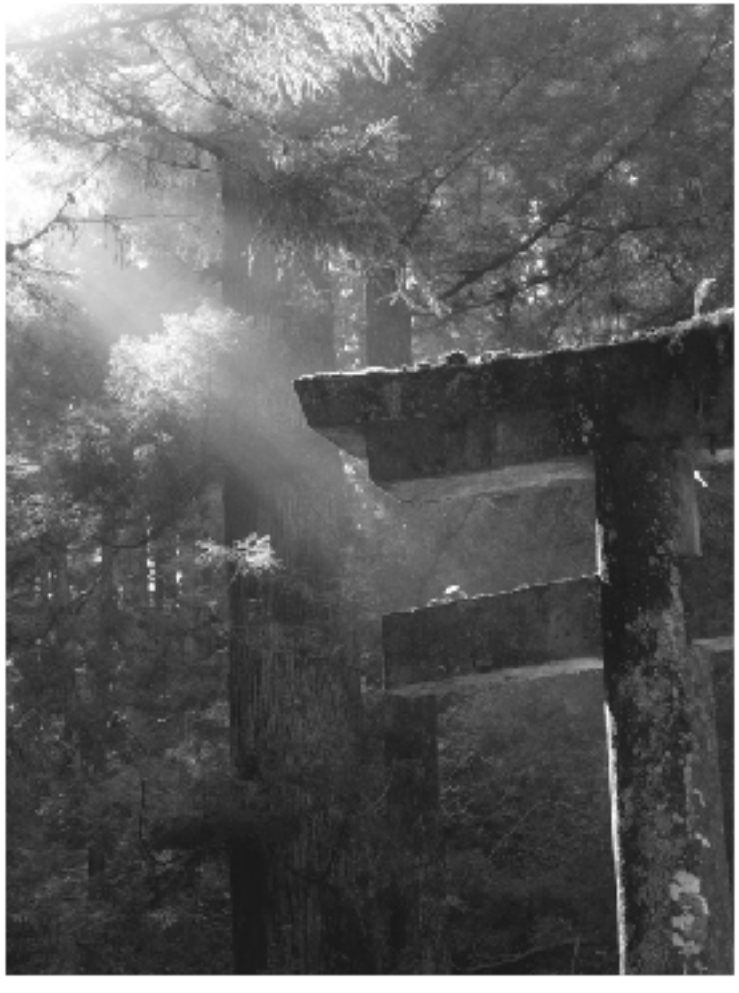


Neural Architectures for Fine-grained Entity Type Classification

Sonse Shimaoka, Pontus Stenetorp,
Kentaro Inui, and Sebastian Riedel

Background

Entity type classification

The game was won by Valencia CF, coached by Salvador González

organization person

Fine-grained entity type classification

The game was won by Valencia CF, coached by Salvador González

Valencia CF	Salvador González
/organization	/person
/organization/sports_club	/person/coach

Terminology

left context right context

The game was won by Valencia CF, coached by Salvador González.

mention

Why fine-grained entity type classification?

- Question answering (Lee et al., 2006)
- Knowledge base population (Carlson et al., 2010)
- Relation extraction (Ling and Weld, 2012)

Previous work

- Ling and Weld (2012)
  - Using distant supervision (Mintz et al., 2009)
- Gillick et al. (2014)
  - Context-dependent
- Yogatama et al. (2015)
  - Embedding-based
- Ren et al. (2016)
  - Data de-noising

Research questions

1. Learned and *hand-crafted* features
2. Exploiting the label *hierarchy*
3. Training data *discrepancies*
4. *Attention analysis*

Data

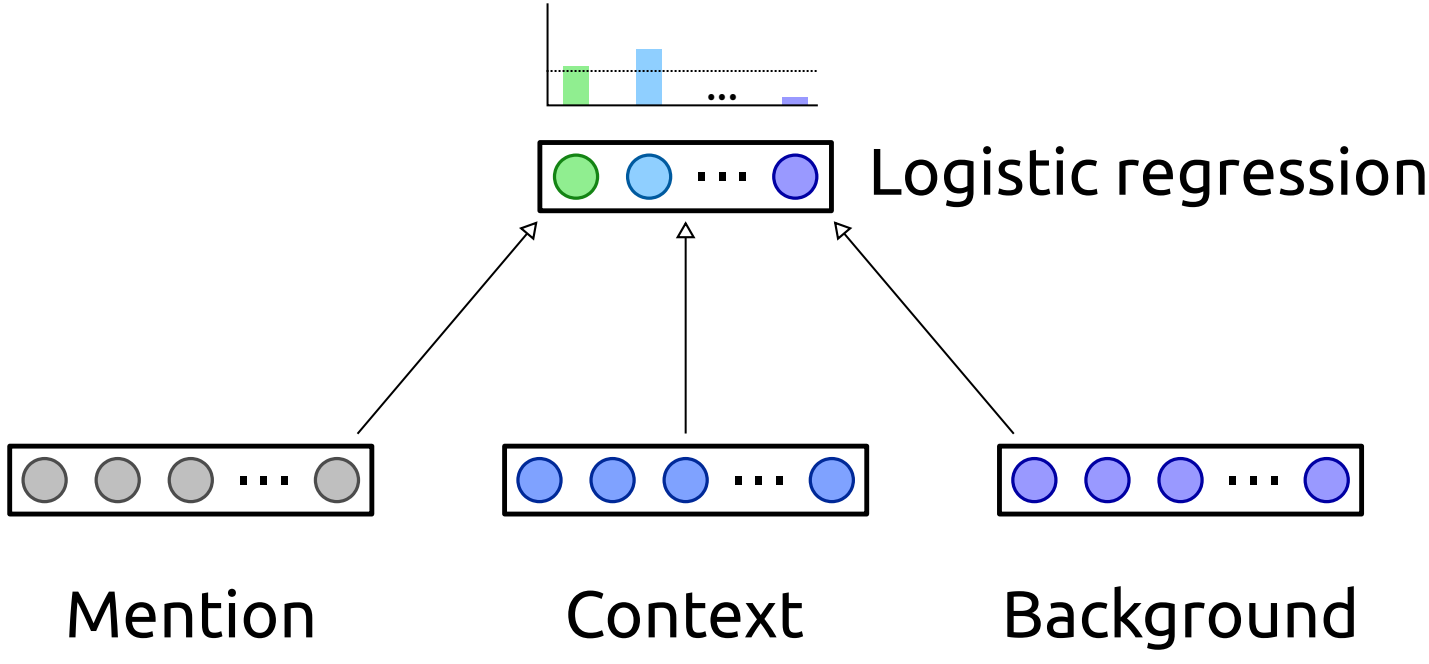
- Training: Ren et al. (2016)
- Evaluation:
  1. FIGER (GOLD) (Ling and Weld, 2012)
  2. OntoNotes (Gillick et al., 2014)

Metrics

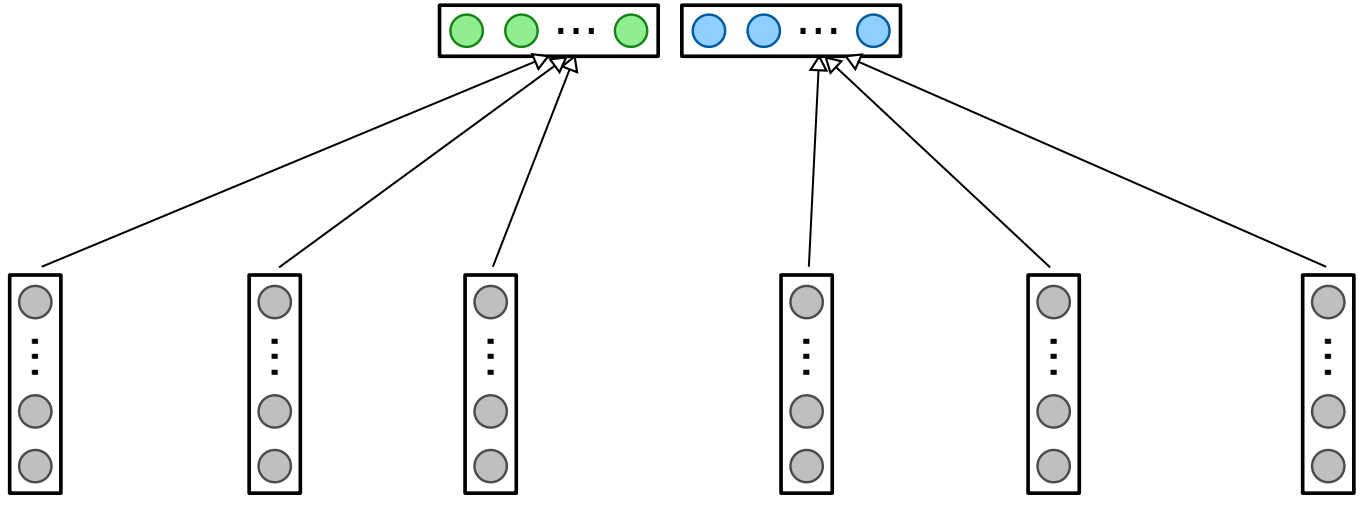
1. Accuracy
2. Loose Macro F1
3. Loose Micro F1

- Same as Ling and Weld (2012).

General model structure

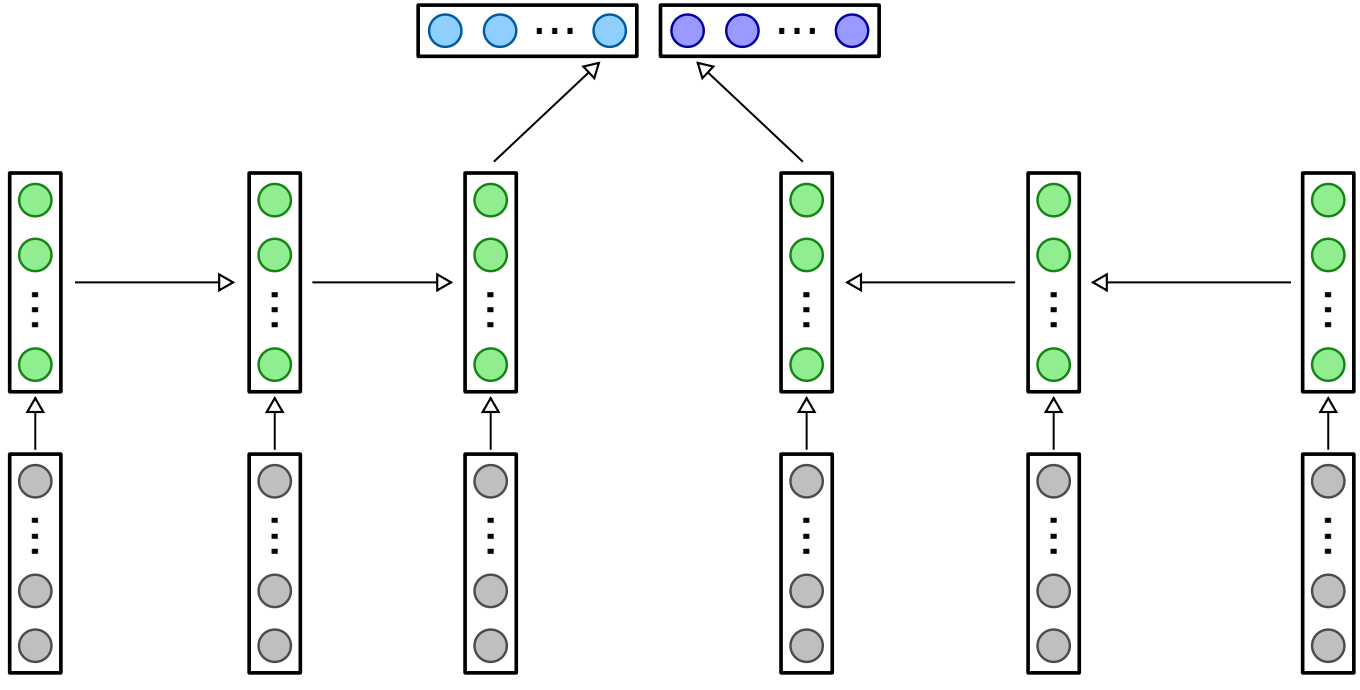


Context representation: Averaging



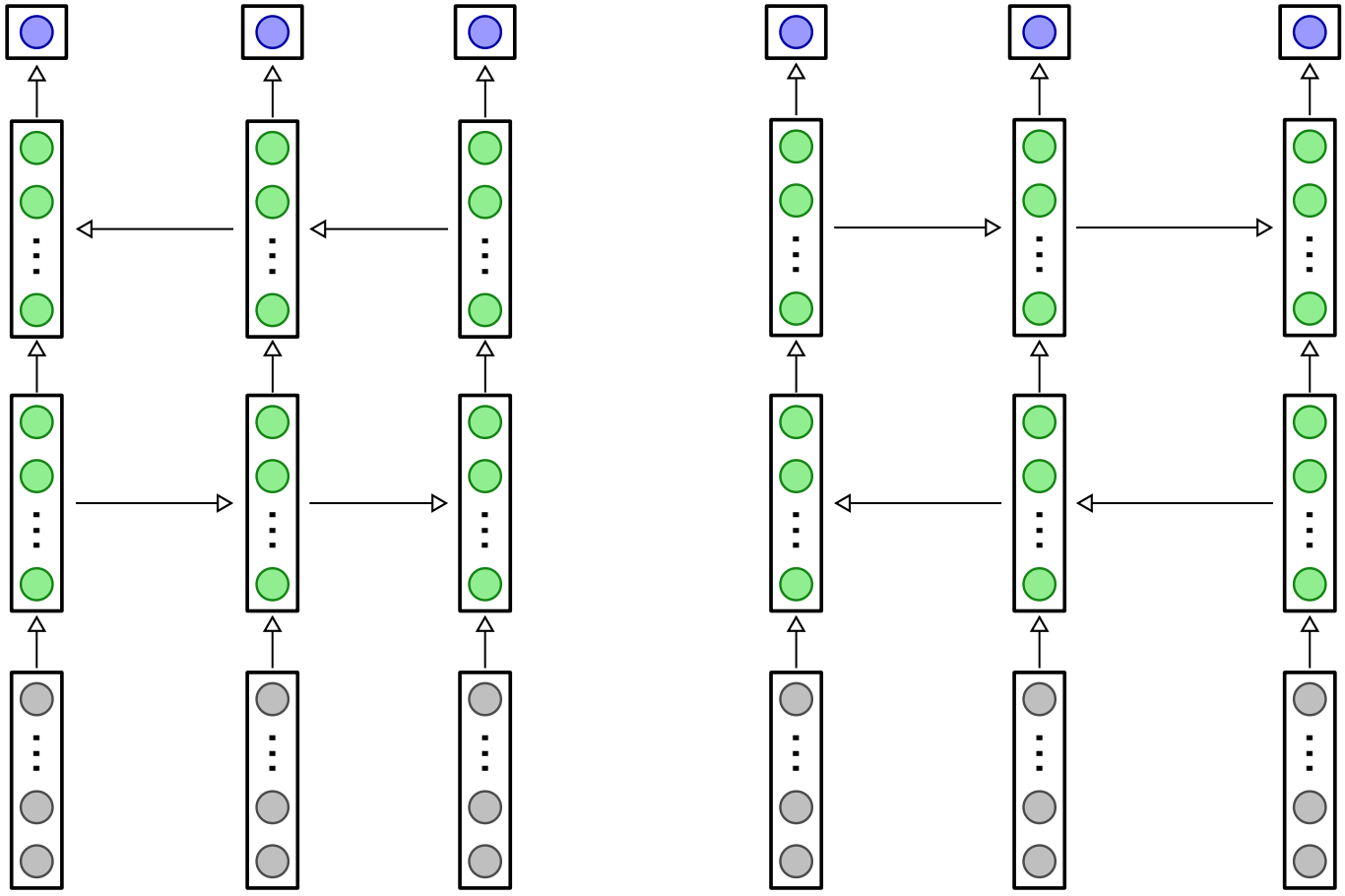
... was won by , coached by ...

Context representation: LSTM



... was won by , coached by ...

Context representation: Attentive



... was won by , coached by ...

Model parameters

- Tuned on development set
- Threshold: 0.5
- Context size: 10
- Embedding size: 200
- LSTM sizes: 100
- Batches of 1,000 using Adam for 5 epochs.
- Dropout: 0.5

Handcrafted features

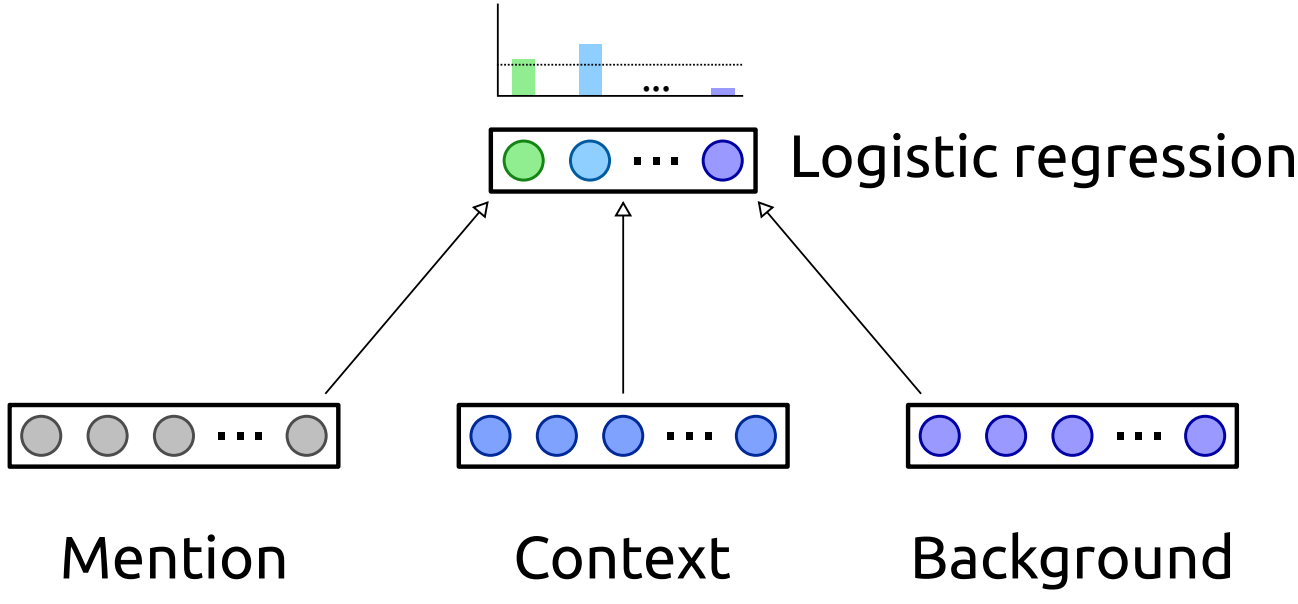
Hand-crafted features

"… who [Barack H. Obama] first picked …"

Feature	Description	Example
Head	Head of mention	Obama
Non-head	Non-head mention tokens	Barack, H.
Cluster	Brown clusters of head	1110, …
Characters	Head character trigrams	:ob, oba, …
Shape	Token shape of mention	Aa A. Aa
Role	Dependency label of the mention head	subj
Context	Tokens before and after mention	B:who, A:first
Parent	The head's lexical parent	picked
Topic	Document LDA topic	LDA:13

Based on features from Gillick et al. (2014)

Hybrid model



Experiment

- Add hand-crafted features to our models.

Results on Figer (GOLD)

Model	Acc.	Macro	Micro
Hand-crafted	51.33	71.91	68.78
Averaging	46.36	71.03	65.31
Averaging + Hand-crafted	52.58	72.33	70.04
LSTM	55.60	75.15	71.73
LSTM + Hand-crafted	57.02	76.98	73.94
Attentive	54.53	74.76	71.58
Attentive + Hand-crafted	**59.68**	**78.97**	**75.36**

Results on OntoNotes

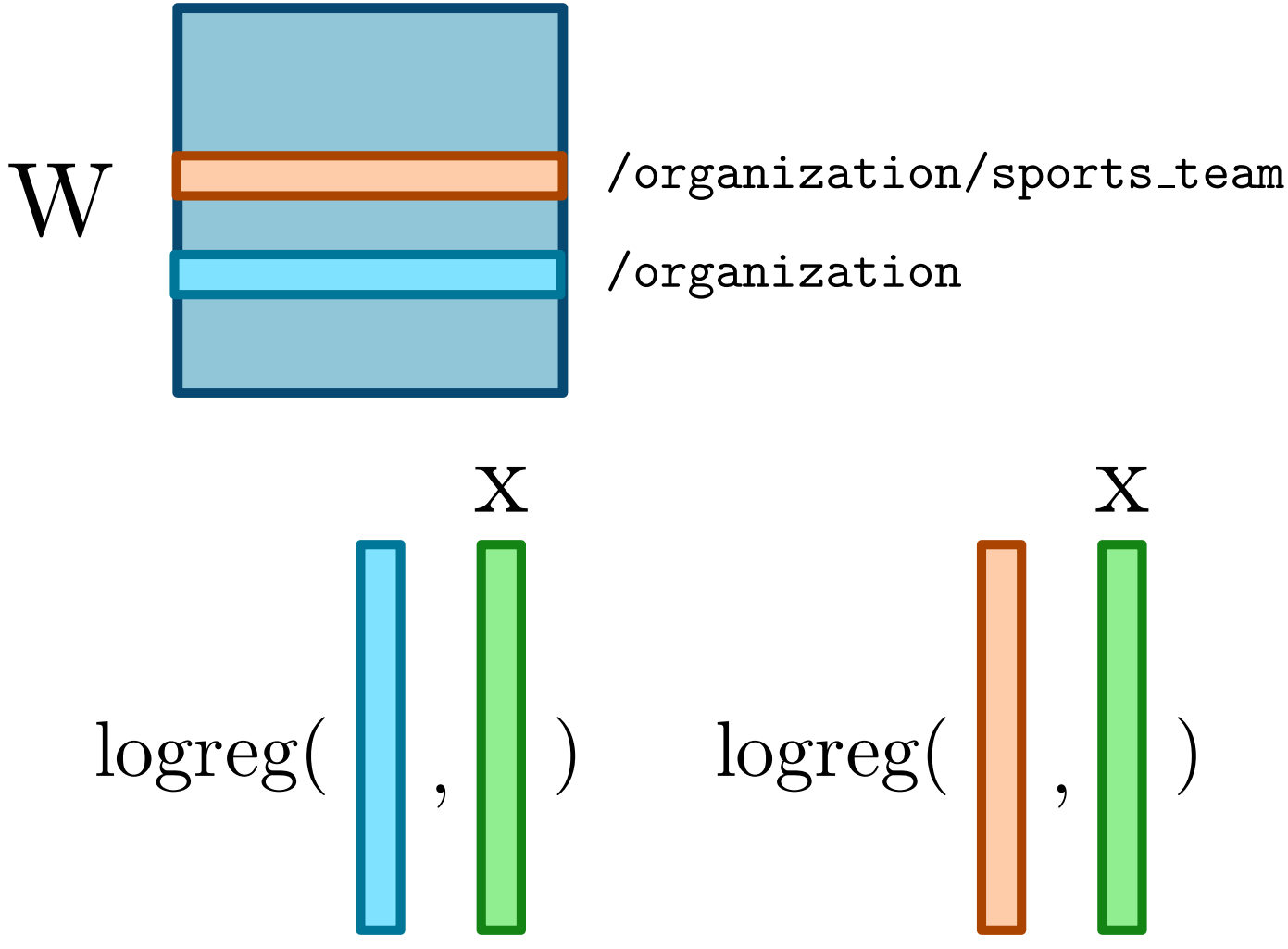
Model	Acc.	Macro	Micro
Hand-crafted	48.16	66.33	60.16
Averaging	46.17	65.26	58.25
Averaging + Hand-crafted	**51.57**	**70.61**	**64.24**
LSTM	49.20	66.72	60.52
LSTM + Hand-crafted	48.58	68.54	62.89
Attentive	50.32	67.95	61.65
Attentive + Hand-crafted	49.54	69.04	63.55

Findings

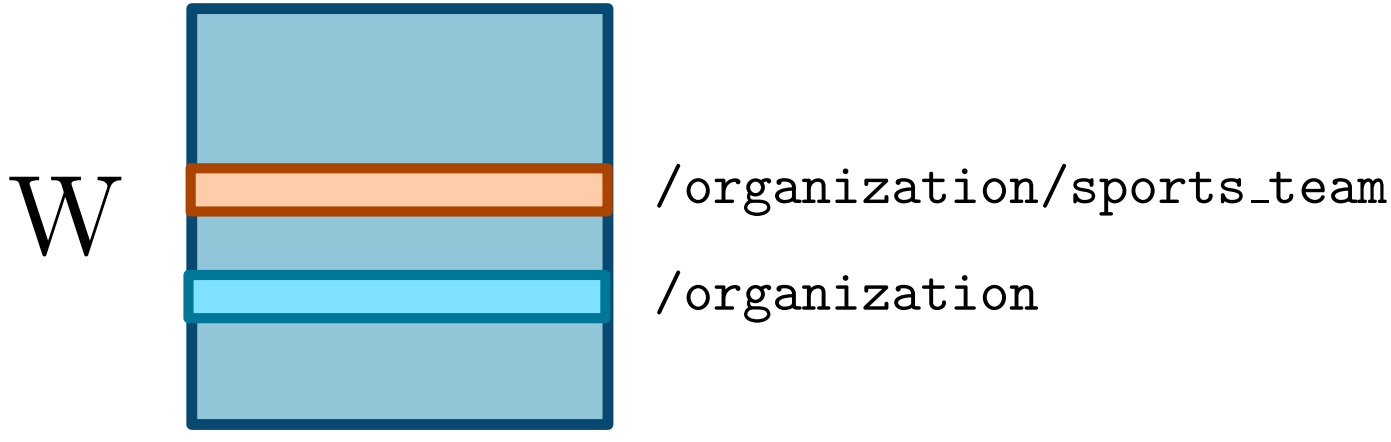
- Consistent increases across both datasets.
- Learnt and hand-crafted complement each other.
- First to consider hand-crafted and attention?

Label hierarchy

Label encoding



Hierarchical label encoding



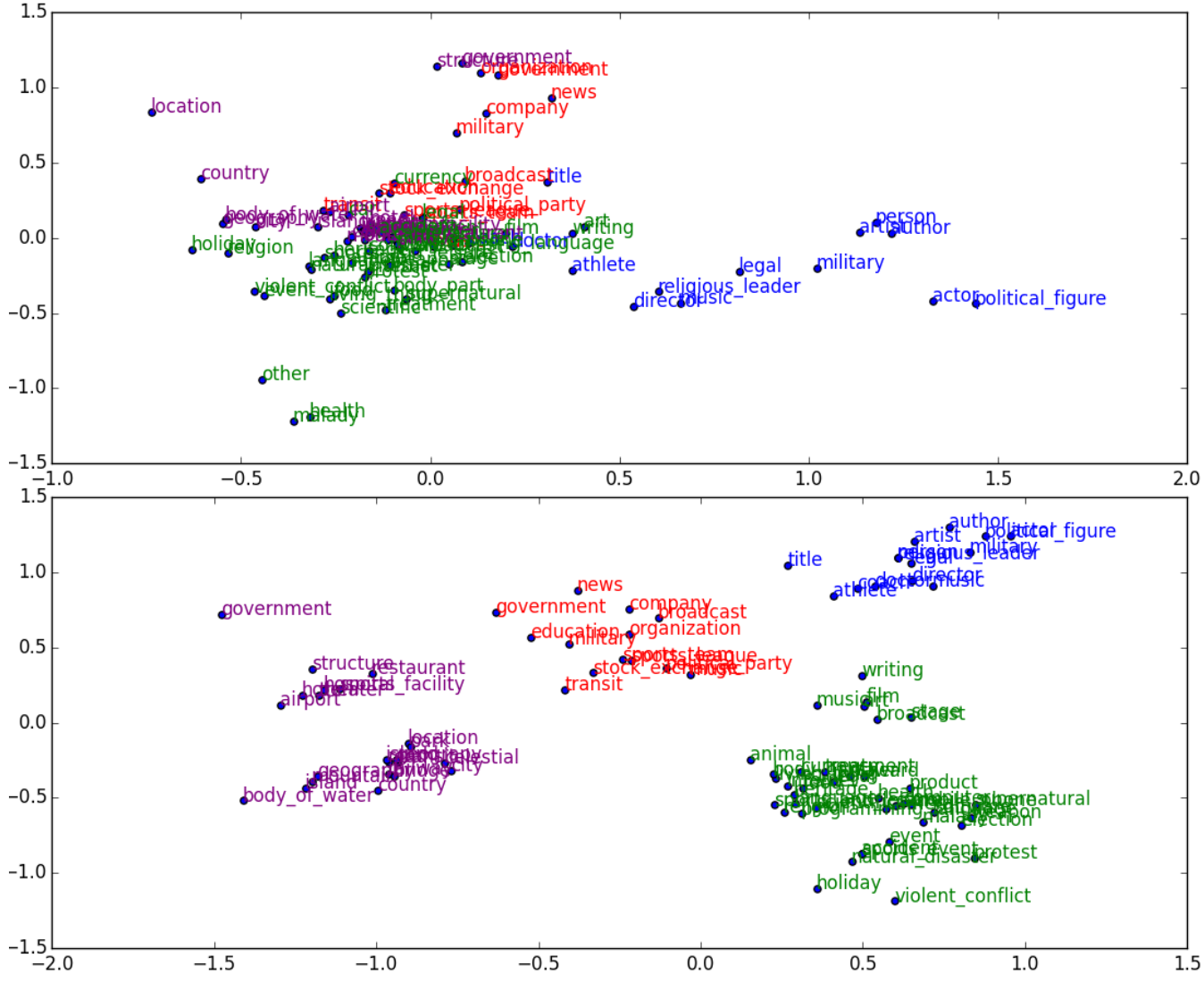
Experiment

- Add hierarchical label encoding to our models.

Results on OntoNotes

Model	Acc.	Macro	Micro
Averaging	46.17	65.26	58.25
Averaging + Hier	47.15	65.53	58.25
Averaging + Hand-crafted	51.57	70.61	64.24
Averaging + Hand-crafted + Hier	**51.74**	**70.98**	64.91
LSTM	49.20	66.72	60.52
LSTM + Hier	48.96	66.51	60.70
LSTM + Hand-crafted	48.58	68.54	62.89
LSTM + Hand-crafted + Hier	50.42	69.99	64.57
Attentive	50.32	67.95	61.65
Attentive + Hier	51.10	68.19	61.57
Attentive + Hand-crafted	49.54	69.04	63.55
Attentive + Hand-crafted + Hier	50.89	70.80	**64.93**

Hierarchical label weights



Findings

- Inconsistent for Figer (Gold)
- Consistent for OntoNotes
- Leads to similar weights for similar labels.

Dataset discrepancies

Evaluation data

1. FIGER (GOLD) (Ling and Weld, 2012)
2. OntoNotes (Gillick et al., 2014)

Training data

1. W2M: 2,000,000 mentions from Wikipedia.
2. W2M+D: Same as W2M, but denoised.
3. W2.6M: Additional 600,000 mentions.
4. GN1: Mentions from Google News.
5. GN2: Mentions from Google News.

Training data divergence

Work	W2M	W2M+D	W2.6M	GN1	GN2
Ling and Weld (2012)	✓				
Gillick et al. (2014)				✗	
Yogatama et al. (2015)					✗
Ren et al. (2016)	✓ *	✓ *			
Shimaoka et al. (2016)			✓		

Experiment

- We have two previously implemented models.
- Performance effect from training data?
- State of the art comparison on unequal footing.

Different training data on Figer (GOLD)

Model	Data	Acc.	Macro	Micro
Attentive (Shimaoka et al., 2016)	W2.6M	58.97	77.96	74.94
Attentive	W2M	54.53	74.76	71.58
Attentive + Hand-crafted	W2M	59.68	**78.97**	**75.36**
Figer + PLE (Ren et al., 2016)	W2M+D	**59.9**	76.3	74.9

Different training data on OntoNotes

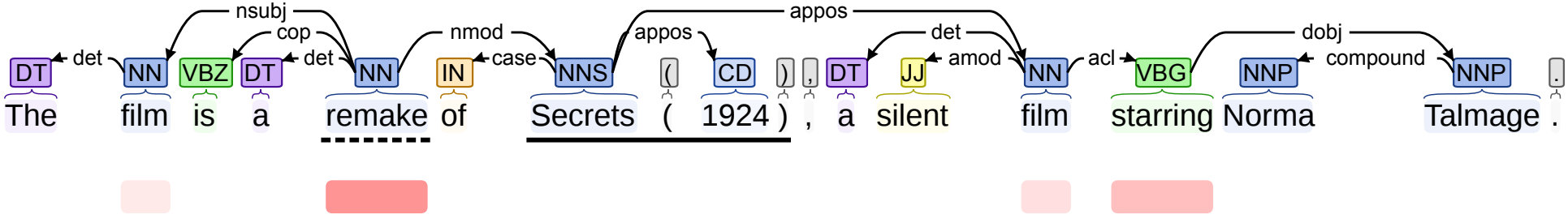
Model	Data	Acc.	Macro	Micro
Hand-crafted (Gillick et al., 2014)	GN1	n/a	n/a	**70.01**
Hand-crafted	W2M	48.16	66.33	60.16
Attentive + Hand-crafted + Hier	W2M	50.89	70.80	64.93
Figer + PLE (Ren et al., 2016)	W2M+D	**57.2**	**71.5**	66.1

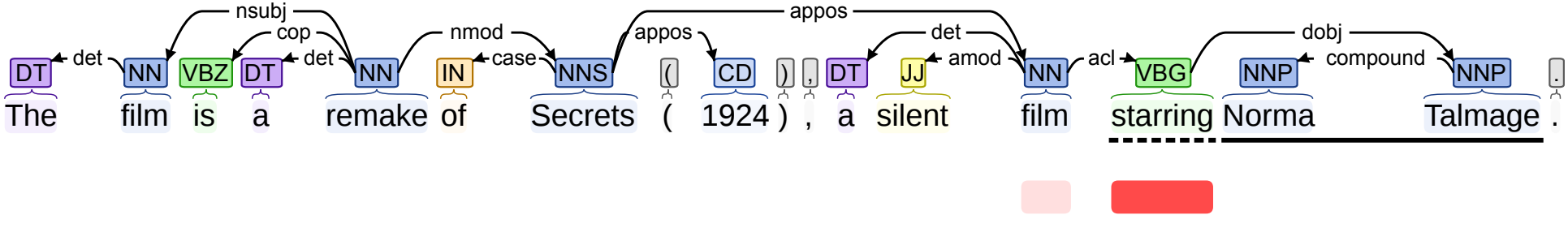
Findings

- Training data has **significant** impact:
  - Attentive: -3.36% Loose Micro F1
  - Hand-crafted: -9.85% Loose Micro F1
- Can we trust previously published numbers?
- State of the art on Figer (Gold), **despite** discrepancy.

Attention analysis

What does the model focus on?





Experiment

- Parse using the Stanford Parser.
- Correlate predictions with mention parent.
- Same as the hand-crafted *Parent* feature.

Attention analysis

Type	Parent	Frequent words
/location	0.319	in, at, born
/organization	0.324	at, the, by
/art/film	0.207	film, films, in
/music	0.259	album, song, single
/award	0.583	won, a, received
/event	0.310	in, during, at

Findings

- Attention focus correlates with mention parent.
- Implicitly learns head finding?
- Explains less benefit from hand-crafted features?

Conclusions and future work

Conclusions

- Learnt and hand-crafted are complementary.
  - Even with attention.
- Inconsistent results for label hierarchy.
  - But clusters label weights.
- Choice of training data has significant impact.
  - Up to 9.85% Loose Micro F1.
- Attention mechanism focuses on mention parent.
  - Implicitly learning head finding?
- State of the art on Figer (Gold), despite discrepancy.
  - Attentive + Hand-crafted: 75.36 Loose Micro F1.

Future work

- Conditioned encoding (Augenstein et al., 2016)
- Further re-implementation of previous models.
- What other linguistic phenomena does the attention learn? (e.g. Kuncoro et al., (2017))

Thank you for your attention

ご清聴ありがとうございました

Tack för er uppmärksamhet

https://github.com/shimaokasonse/NFGEC

Same training data on Figer (GOLD)

Model	Acc.	Macro	Micro
Hand-crafted	51.33	71.91	68.78
Averaging	46.36	71.03	65.31
Averaging + Hand-crafted	52.58	72.33	70.04
LSTM	55.60	75.15	71.73
LSTM + Hand-crafted	57.02	76.98	73.94
Attentive	54.53	74.76	71.58
Attentive + Hand-crafted	59.68	78.97	75.36
Figer (Ling and Weld, 2012)	52.30	69.90	69.30
Figer (Ren et al., 2016)	47.4	69.2	65.5

Different training data on Figer (GOLD)

Model	Data	Acc.	Macro	Micro
Attentive + Hand-crafted	W2M	59.68	78.97	75.36
Attentive (Shimaoka et al., 2016)	W2.6M	58.97	77.96	74.94
Figer + PLE (Ren et al., 2016)	W2M+D	59.9	76.3	74.9
HYENA + PLE (Ren et al., 2016)	W2M+D	54.2	69.5	68.1
K-WASABIE (Yogatama et al., 2015)	GN2	n/a	n/a	72.25

Same training data on OntoNotes

Model	Acc.	Macro	Micro
Hand-crafted	48.16	66.33	60.16
Averaging	46.17	65.26	58.25
Averaging + Hier	47.15	65.53	58.25
Averaging + Hand-crafted	51.57	70.61	64.24
Averaging + Hand-crafted + Hier	51.74	70.98	64.91
LSTM	49.20	66.72	60.52
LSTM + Hier	48.96	66.51	60.70
LSTM + Hand-crafted	48.58	68.54	62.89
LSTM + Hand-crafted + Hier	50.42	69.99	64.57
Attentive	50.32	67.95	61.65
Attentive + Hier	51.10	68.19	61.57
Attentive + Hand-crafted	49.54	69.04	63.55
Attentive + Hand-crafted + Hier	50.89	70.80	64.93
Figer (Ren et al., 2016)	36.90	57.80	51.60

Different training data on OntoNotes

Model	Data	Acc.	Macro	Micro
Averaging + Hand-crafted + Hier	W2M	51.74	70.98	64.91
Attentive + Hand-crafted + Hier	W2M	50.89	70.80	64.93
Figer + PLE (Ren et al., 2016)	W2M+D	57.2	71.5	66.1
HYENA + PLE (Ren et al., 2016)	W2M+D	54.6	69.2	62.5
Hand-crafted (Gillick et al., 2014)	GN1	n/a	n/a	70.01
K-WASABIE (Yogatama et al., 2015)	GN2	n/a	n/a	72.98