Hebrew NER

Named Entity Recognition

Dan Bareket
ONLP & OMILAB

ONLP Meetup, April 2019

Task

- **Input**: *text*
- **Output**: *named entity mentions*
- Every mention includes:
  - *Borders*
  - *Category*
- Strict evaluation - exact entities (border and class)
- In Hebrew

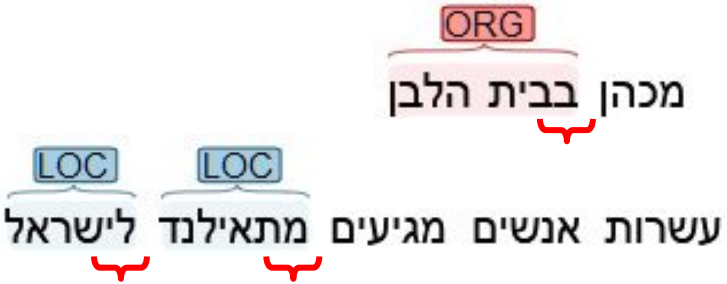
שבוע ה בא ידון מנכ"ל התאחדות ה איכרים , שלמה רייזמן עם מנכ"ל שירות ה תעסוקה , דוד מנע

Task

- **Input**: *text*
- **Output**: *named entity mentions*
- Every mention includes:
  - *Borders*
  - *Category*
- Strict evaluation - exact entities (border and class)
- In Hebrew

Previous Hebrew NER work

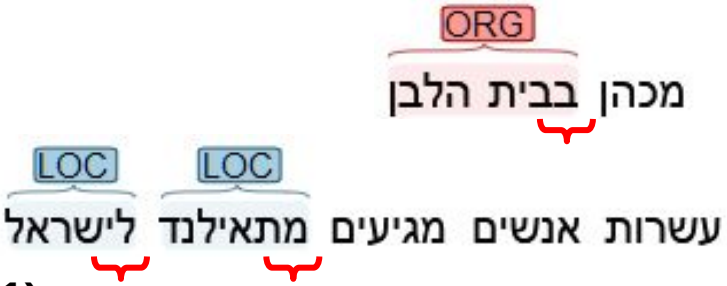
- Mordecai and Elhadad, 2005
- Corpus:
  - ~57k tokens, ~4700 NE mentions
  - News: 7 *ערוץ, הארץ, מעריב*
  - Guidelines based on MUC7 ('98) & CoNLL 2003
- **No morphology** or token segmentation



- **Classical models**, feature engineering
  - Regex+Lexicon
  - HMM
  - MEMM
  - Combined

Previous Hebrew NER work

- Mordecai and Elhadad, 2005
- Corpus:
  - ~57k tokens, ~4700 NE mentions
  - News: 7 *ערוץ, הארץ, מעריב*
  - Guidelines based on MUC7 ('98) & CoNLL 2003
- **No morphology** or token segmentation



- **Classical models**, feature engineering **(F1)**
  - Regex+Lexicon **(58)**
  - HMM **(68)**
  - MEMM **(76)**
  - Combined **(79)**

Hebrew NER - A new research agenda

- Moving from surface tokens to morphemes
  - Token / Word / Char based
  - Morpheme based ← YAP!
- Moving from classical ML to neural
- External signals in a neural morpheme world?
  - Linguistic knowledge - POS, morphology, etc. ← YAP!
  - “World” knowledge - pre-trained embeddings ← YAP!
  - Task-specific knowledge - gazetteers, lexicons, WikiData etc.

Hebrew NER - A new research agenda

- Moving from surface tokens to morphemes
  - Token / Word / Char based
  - Morpheme based ← YAP!
- Moving from classical ML to neural
- External signals in a neural morpheme world?
  - Linguistic knowledge - POS, morphology, etc. ← YAP!
  - “World” knowledge - pre-trained embeddings ← YAP!
  - Task-specific knowledge - gazetteers, lexicons, WikiData etc.
- Two paths:
  - **New corpus** with “gold” morphology & NER → Hebrew Treebank
  - Strong neural **baseline models**

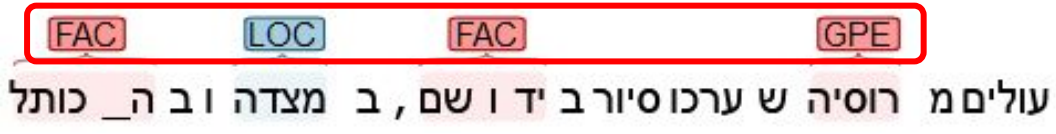
New Corpus

- **Morphology**

New Corpus

- Morphology

- **<u>Fine-grained entity types (as in OntoNotes 5.0)</u>**



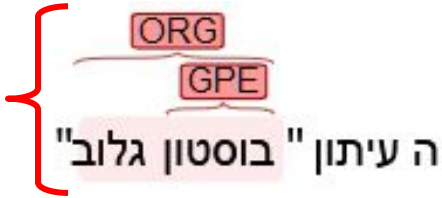
2005	Ours
Person	Person
Location	Geo-Political-Entity
	Location
	Facility
Organization	Organization
Misc-Event	Event
Misc-Affiliation	___
Misc	Work-of-Art
	Product
	Language

New Corpus

- Morphology

- Fine-grained entity types (as in OntoNotes 5.0)

- **<u>Nesting</u>**



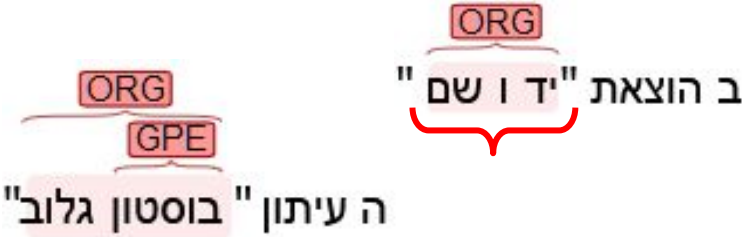
2005	Ours
Person	Person
Location	Geo-Political-Entity
	Location
	Facility
Organization	Organization
Misc-Event	Event
Misc-Affiliation	
Misc	Work-of-Art
	Product
	Language

New Corpus

- Morphology

- Fine-grained entity types (as in OntoNotes 5.0)

- Nesting



- **<u>Reference, not surface</u>**

2005	Ours
Person	Person
Location	Geo-Political-Entity
	Location
	Facility
Organization	Organization
Misc-Event	Event
Misc-Affiliation	
Misc	Work-of-Art
	Product
	Language

Annotation

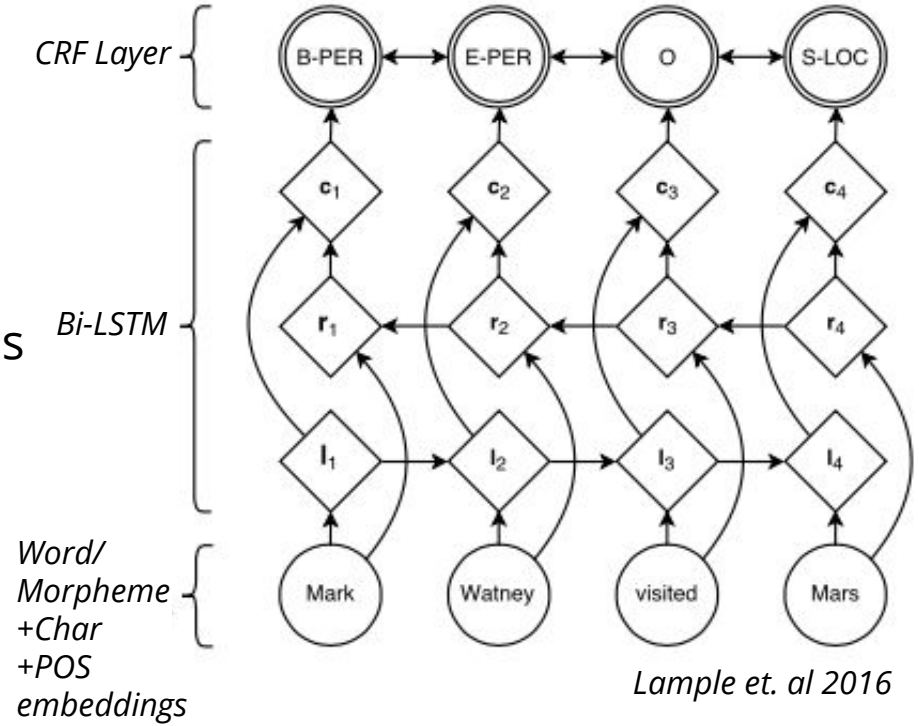
- **Hebrew Treebank** - *Ha'aretz*, ~6200 Sentences, ~161k tokens
- **WebAnno** - **<u>GREAT TOOL!</u>** (Technische Universität Darmstadt)
- Annotation process
  - 2-step pilot (12 participants)
  - Guidelines ver. 1
  - Whole corpus - 2 annotators - w/ incremental guideline updates
  - Curation - disagreements and backward fixes
- 7800~ bottom layer mentions, 1100~ more in nested mentions

Annotation

- **Hebrew Treebank** - *Ha'aretz*, ~6200 Sentences, ~161k tokens
- **WebAnno** - **<u>GREAT TOOL!</u>** (Technische Universität Darmstadt)
- Annotation process
  - 2-step pilot (12 participants)
  - Guidelines ver. 1
  - Whole corpus - 2 annotators - w/ incremental guideline updates
  - Curation - disagreements and backward fixes
- 7800~ bottom layer mentions, 1100~ more in nested mentions
- Agreement (F1, bottom layer)
  - Inter-annotator Agreement (**89**)
  - A with Curation (**96**)
  - B with Curation (**92**)

Baseline neural model architecture

- **Bi-LSTM** encoder
- Input layer is either:
  - **Original words**
  - **Morphemes** ← YAP
- **CRF** output layer
- Extra word/morpheme level features
  - **Pre-trained fastText** embeddings
  - **POS** ← YAP
  - **Character embeddings** (Bi-LSTM)



Lample et. al 2016

(Preliminary) F1 scores - Mordecai corpus

Bi-LSTM+CRF with random initial embeddings

	Clean	***+POS***	***+Char***	***+POS+Char***
Word	51.9	61.9	60.2	65.4
Morpheme	**55.2**	**64.3**	**65.7**	**67.6**

*Bi-LSTM+CRF with **fastText** initial embeddings*

	fastText	***+POS***	***+Char***	***+POS+Char***
Word	**73.5**	74.5	77.2	78.5
Morpheme	72.5	**75.2**	**79.0**	**80.2**

(Preliminary) F1 scores - Word→Morph

Bi-LSTM+CRF with random initial embeddings

	Clean	***+POS***	***+Char***	***+POS+Char***
Word	51.9	61.9	60.2	65.4
Morpheme	(+6%) **55.2**	(+4%) **64.3**	(+9%) **65.7**	(+3%) **67.6**

*Bi-LSTM+CRF with **fastText** initial embeddings*

	fastText	***+POS***	***+Char***	***+POS+Char***
Word	**73.5**	74.5	77.2	78.5
Morpheme	(-1%) 72.5	(+1%) **75.2**	(+2%) **79.0**	(+2%) **80.2**

(Preliminary) F1 scores - Adding fastText

Bi-LSTM+CRF with random initial embeddings

	Clean	***+POS***	***+Char***	***+POS+Char***
Word	51.9	61.9	60.2	65.4
Morpheme	**55.2**	**64.3**	**65.7**	**67.6**

*Bi-LSTM+CRF with **fastText** initial embeddings*

	fastText	***+POS***	***+Char***	***+POS+Char***
Word	(+42%) **73.5**	(+20%) 74.5	(+28%) 77.2	(+20%) 78.5
Morpheme	(+31%) 72.5	(+17%) **75.2**	(+20%) **79.0**	(+18%) **80.2**

Conclusions

- Goal - Adding NER to Hebrew NLP pipeline
- A new benchmark corpus for NER with morphology
  - Researching NER ⇆ Morphology interaction
  - NOT JUST ABOUT SCORES, better captures phenomena
- YAP enables new research directions
  - E.g. morpheme RNNs, training morphological embeddings
- We're working on multiple future directions - stay tuned!
  - Joint morphology + NER?
  - Morphological embeddings?
  - Incorporating task-specific data (gazetteers, WikiData)
  - and more...