How to Use Gazetteers for Entity Recognition with Neural Models

Simone Magnolini¹, Valerio Piccioni², Vevake Balaraman¹,², Marco Guerini¹, Bernardo Magnini¹

¹ Fondazione Bruno Kessler, Trento, Italy
² University of Trento

magnini@fbk.eu
Entity Recognition (1)

| Prof. | B. | Magnini | teaches | Computational | Linguistics | at | University | of | Trento |
|-------|----|---------|---------|---------------|-------------|----|------------|----|--------|
| O     | B-PER | I-PER | O       | O             | O           | O  | B-ORG     | I-ORG | I-ORG |

- **Named Entities (NER)**
  - Proper nouns: almost fixed expressions (e.g. no morphological variations)
  - Common categories: PERSON, LOCATION, ORGANIZATION
  - Several datasets available (e.g. CoNLL, OntoNotes)
Entity Recognition (2)

|   | I  | would | like | a   | salami | pizza | and | two | cheese | sandwiches |
|---|----|-------|------|-----|--------|-------|-----|-----|--------|------------|
| O | O  | O     | O    | O   | B-FOOD | I-FOOD| O   | O   | B-FOOD | I-FOOD     |

- **Nominal Entities**
  - Noun phrases: compositional
    - *pasta*
    - *pasta with pesto*: + prepositional modifier
    - *Italian pasta with pesto*: + adjectival modifier
    - *spaghetti with broccoli:*
      - *spaghetti with pesto*: can be inferred
  - Some categories: FOOD, FORNITURE, CLOTHES, etc.
  - Few datasets available
Entity Recognition (3)

- **Neural models**
  - State of art performance
  - Data-driven (typically few thousand labeled sentences)
  - No need for hand-crafted features (e.g. capitalization)
  - No need for external knowledge sources

NeuroNLP2: Ma and Hovy, ACL 2016
Using Gazetteers with Neural Models

- Gazetteers
  - Lists of entity names for a certain category
  - Rich source of domain knowledge
  - Relatively cheap to obtain in large quantities (100K+)

- PERSON
  - Tom Cruise
  - Nicolas Cage
  - Bruce Willis
  - John Wayne
  - Christian Bale
  - Bernardo Magnini

- FOOD
  - spaghetti al pesto
  - lasagne
  - pasta
  - mozzarella cheese
  - hot smoked salmon
  - roasted salmon
  - sandwich
Using Gazetteers with Neural Models

Research questions

• Is domain knowledge (gazetteers) useful with neural models of entity labeling?
• How can we inject such knowledge?
• Named entities vs nominal entities
• Impact of the dataset size

Related work
Liu et al: Towards Improving Neural Named Entity Recognition with Gazetteers - ACL 2019
Experimental Method

1. Extract gazetteer features, three methods:
   - Single token presence
   - Multi-token presence
   - Neural model of the gazetteer

2. Chose a neural sequence labeling system
   - NeuroNLP2 (Ma and Hovy, ACL 2016) is used in this study

3. Integrate gazetteer features into the neural labeling system
   - As additional embedding dimension
   - As features for the classifier
Bruce Wayne (Christian Bale) is attacked during a cruise by his nemesis Bane that throws him in a cage...
Gazetteer as feature: Single Token Presence

Bruce Wayne (Christian Bale) is attacked during a cruise by his nemesis Bane (Tom Hardy) that throws him in a cage…

A token has a positive value if it matches with a token present in the gazetteer.

Feature: 1-hot vector for each class to be labeled
Gazetteer as feature: Multi-token Presence

Bruce Wayne (Christian Bale) is attacked during a cruise by his nemesis Bane (Tom Hardy) that throws him in a cage…

Bruce Wayne (Christian Bale) is attacked during a cruise by his nemesis Bane (Tom Hardy) that throws him in a cage…

A token has a positive value if it is part of an entity name in the gazetteer.

Feature: 1-hot vector for each class to be labeled
Gazetteer as feature: Gazetteer Neural Model - $\text{NN}_g$

Bruce Wayne (Christian Bale) is attacked during a cruise by his nemesis Bane (Tom Hardy) that throws him in a cage…

A token has a positive value if it is classified as part of a category entity by a gazetteer classifier ($\text{NN}_g$).

- $\text{NN}_g$ is pre-trained on the Gazetteer (Guerini et al. SigDial 2018)
- The representation of the input sequence at the last layer before Softmax is used as the corresponding gazetteer representation for the sequence.

Feature: 1-hot vector for each class to be labeled
Reference Sequence Labeling Model: NeuroNLP2

- **NeuroNLP2**: Ma and Hovy, ACL 2016 – code available

- **Embedding**
  - Character → CNN (30-d)
  - Word → GloVe (100-d)

- **Recurrent NN**
  - BiLSTM

- **Output layer**
  - CRF

---

NeuroNLP2: Ma and Hovy, ACL 2016
Integration 1: Enriching Embeddings

\[ \text{Embedding}(x) = [x_{\text{word}}; x_{\text{char}}] \]

\[ \text{Embedding}(x) = [x_{\text{word}}; x_{\text{char}}; x_{\text{gaz}}] \]
Integration 2: CRF classifier

\[ \text{hidden}_x = [LSTM_x; LSTM_x] \]

\[ \text{hidden}_x = [LSTM_x; LSTM_x; GAZ_x] \]
## Experimental Setting: Two Datasets

| Dataset                        | Token | Types | Entities | Sentences |
|-------------------------------|-------|-------|----------|-----------|
| **CoNLL-2003 – named entities (English, entities=23K)** |       |       |          |           |
| Train                         | 204567| 23624 | 23499    | 14987     |
| Dev                           | 51578 | 9967  | 5942     | 3466      |
| Test                          | 46666 | 9489  | 5648     | 3684      |
| **Diabetic Patient Diary (DPD) – nominal entities (food, Italian, entities=1,7K)** |       |       |          |           |
| Train                         | 4748  | 636   | 1757     | 450       |
| Dev                           | 296   | 138   | 122      | 49        |
| Test                          | 2315  | 379   | 583      | 200       |
### Experimental Setting: Gazetteers

| entity | #entities | #tokens | length± SD | TTR | Type1 (%) | Type2 (%) | Sub-entity (%) |
|--------|-----------|---------|------------|-----|-----------|-----------|---------------|
| **CoNLL gazetteers** | | | | | | | |
| PER    | 3613      | 6454    | 1.79±0.54  | 0.96| 0.99      | 0.63      | 23.60         |
| LOC    | 1331      | 1720    | 1.29±0.69  | 0.97| 0.97      | 0.66      | 10.14         |
| ORG    | 2401      | 4659    | 1.94±1.44  | 0.91| 0.91      | 0.35      | 19.44         |
| MISC   | 869       | 1422    | 1.14±0.94  | 0.89| 0.89      | 0.73      | 19.85         |
| **DPD gazetteer** | | | | | | | |
| FOOD   | 23472     | 83264   | 3.55±1.87  | 0.75| 0.75      | 1.22      | 11.27         |

Food gazetteer is much bigger than CoNLL gazetteers
### Experimental Setting: Gazetteers

| Entity   | #entities | #tokens | length± SD | TTR | Type1 (%) | Type2 (%) | Sub-entity (%) |
|----------|-----------|---------|------------|-----|-----------|-----------|----------------|
| **CoNLL gazetteers** | | | | | | | |
| PER      | 3613      | 6454    | 1.79±0.54  | 0.96| 19.07     | 04.63     | 23.60          |
| LOC      | 1331      | 1720    | 1.29±0.69  | 0.97| 04.66     | 10.14     |                |
| ORG      | 2401      | 4659    | 1.94±1.16  | 0.91| 04.33     | 15.06     | 19.44          |
| MISC     | 869       | 1422    | 1.64±0.94  | 0.85| 08.61     | 08.73     | 19.85          |
| **DPD gazetteer** | | | | | | | |
| FOOD     | 23472     | 83264   | 3.55±1.87  | 0.75| 17.22     | 22.97     | 11.27          |

Food names have much higher length than CoNLL names.
## Results: Gazetteers Integrated at Embedding Level

|               | CoNLL          | DPD            |
|---------------|----------------|----------------|
|               | Accuracy | Precision | Recall | F1    | Accuracy | Precision | Recall | F1    |
| NeuroNL P2    | 98.06    | 91.42      | 90.95  | 91.19 | 88.47    | 77.17      | 74.79  | 75.96 |
|               | + single token | 98.06  | 91.53  | 90.51  | 91.02 | 88.29    | 75.63      | 77.19  | 76.40 |
|               | + multi token | 98.08  | 91.41  | 90.76  | 91.08 | 88.98    | 78.90      | 76.33  | 77.59 |
| + NN\_g       | 98.05    | 91.41      | 91.02  | **91.22** | 89.89    | 79.68      | 77.36  | **78.50** |

Adding gazetteer features at the embedding level works better than CRF integration.
Results: Gazetteers Integrated at Embedding Level

|                | CoNLL          | DPD           |
|----------------|----------------|---------------|
|                | Accuracy       | Precision     | Recall | F1  | Accuracy | Precision | Recall | F1  |
| NeuroNL P2     | 98.06          | 91.42         | 90.95  | 91.19 | 88.47    | 77.17     | 74.79  | 75.96|
| + single token | 98.06          | 91.53         | 90.51  | 91.02 | 88.29    | 75.63     | 77.19  | 76.40|
| + multi token  | 98.08          | 91.41         | 90.76  | 91.08 | 88.98    | 78.90     | 76.33  | 77.59|
| + NNg          | 98.05          | 91.41         | 91.02  | **91.22** | 89.89    | 79.68     | 77.36  | **78.50** |

Small improvement with named entities
## Results: Gazetteers Integrated at Embedding Level

| Gazetteers               | CoNLL Accuracy | CoNLL Precision | CoNLL Recall | CoNLL F1  | Dated Fusion Accuracy | Dated Fusion Precision | Dated Fusion Recall | Dated Fusion F1  |
|--------------------------|----------------|-----------------|--------------|-----------|------------------------|------------------------|---------------------|-----------------|
| NeuroNLP                  | 98.06          | 91.42           | 90.95        | 91.19     | 88.47                  | 77.17                  | 74.00               | 75.96           |
| + single token           | 98.06          | 91.53           | 90.51        | 91.02     | 88.29                  | 75.63                  | 77.19               | 76.40           |
| + multi token            | 98.08          | 91.41           | 90.76        | 91.08     | 88.98                  | 78.90                  | 76.33               | 77.59           |
| + NN_{g}                  | 98.05          | 91.41           | 91.02        | **91.22** | 89.89                  | 79.68                  | 77.36               | **78.50**       |

Higher improvement with nominal entities.
Learning Curve: Nominal entities (DPD)

- NNg is more robust and consistent on nominal entities (DPD)
Learning Curve: named entities (CoNLL)

The multi-token approach works better for named entities (CoNLL)
Conclusion

- Gazetteers are still useful for neural models, under certain conditions:
  - When added as additional features with embeddings
  - When training data are limited
  - When nominal entities are addressed

- Best results are obtained with nominal entities and with a neural classifier built on top of the gazetteer