# COMP 3225

# Natural Language Processing

## Named Entity Recognition

Stuart E. Middleton

sem03@soton.ac.uk

University of Southampton

# Overview

- Named Entity Recognition (NER)
- Conditional Random Fields (CRF)
- Feature Sets

- <break - discussion point>

- Inference for NER
- Evaluation of NER
- Deep learning models for NER

# Named Entity Recognition (NER)

- Named Entity is anything referred to with a proper name
  - Person, Location, Organization, Event ...
  - often types are extended to include Date, Time, Money ...
- Named Entity Recognition (NER) is the task of labelling a text span with the types of named entities

Citing high fuel prices, [$_{\text{ORG}}$ **United Airlines**] said [$_{\text{TIME}}$ **Friday**] it has increased fares by [$_{\text{MONEY}}$ **$6**] per round trip on flights to some cities also served by lower-cost carriers. [$_{\text{ORG}}$ **American Airlines**], a unit of [$_{\text{ORG}}$ **AMR Corp.**], immediately matched the move, spokesman [$_{\text{PER}}$ **Tim Wagner**] said. [$_{\text{ORG}}$ **United**], a unit of [$_{\text{ORG}}$ **UAL Corp.**], said the increase took effect [$_{\text{TIME}}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ **Chicago**] to [$_{\text{LOC}}$ **Dallas**] and [$_{\text{LOC}}$ **Denver**] to [$_{\text{LOC}}$ **San Francisco**].

# Named Entity Recognition (NER)

- Named Entity is anything referred to with a proper name
  - Person, Location, Organization, Event ...
  - often types are extended to include Date, Time, Money ...
- Named Entity Recognition (NER) is the task of labelling a text span with the types of named entities

| Type | Tag | Sample Categories | Example sentences |
|---|---|---|---|
| People | PER | people, characters | **Turing** is a giant of computer science. |
| Organization | ORG | companies, sports teams | The **IPCC** warned about the cyclone. |
| Location | LOC | regions, mountains, seas | **Mt. Sanitas** is in **Sunshine Canyon**. |
| Geo-Political Entity | GPE | countries, states | **Palo Alto** is raising the fees for parking. |

# Named Entity Recognition (NER)

- Named Entity is anything referred to with a proper name
  - Person, Location, Organization, Event ...
  - often types are extended to include Date, Time, Money ...
- Named Entity Recognition (NER) is the task of labelling a text span with the types of named entities
- There are different named entity tagsets
  - Automatic Content Extraction (ACE) Program defines 7 types
  - Many tagsets are based on ACE, with domain-specific types added
- Challenges
  - NER works with spans (multiple words in a sequence) of text
  - Named entity type ambiguity

[$_{\text{PER}}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{\text{ORG}}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{\text{LOC}}$ Washington] for what may well be his last state visit.
In June, [$_{\text{GPE}}$ Washington] passed a primary seatbelt law.

# Named Entity Recognition (NER)

B I O

BIOES

- BIO tagging is a common approach for sequence labelling requiring span-recognition
  - Begin, Inside, Outside >> BIO
  - Begin, Inside, Outside, End, Single >> BIOES
- Append NE type to BIO tag to create a token label

| Words | IO Label | BIO Label | BIOES Label |
|---|---|---|---|
| Jane | I-PER | B-PER | B-PER |
| Villanueva | I-PER | I-PER | E-PER |
| of | O | O | O |
| United | I-ORG | B-ORG | B-ORG |
| Airlines | I-ORG | I-ORG | I-ORG |
| Holding | I-ORG | I-ORG | E-ORG |
| discussed | O | O | O |
| the | O | O | O |
| Chicago | I-LOC | B-LOC | S-LOC |
| route | O | O | O |
| . | O | O | O |

# Conditional Random Fields (CRF)

- Models for NER
  - Non-word features such as capitalization are useful for NER
  - Hidden Markov Model (HMM) is generative, and it is hard to add feature patterns (as opposed to concrete words in sequence)
- Conditional Random Fields (CRF)
  - Discriminative sequence model based on a log-linear model (like logistic regression)
  - Widely used for this type of sequence labelling problem
  - We will describe linear chain CRF here
- Sequence labelling task
  - Input >> X >> Sequence of words
  - Output >> Y >> Sequence of BIO tags
  - len(X) == len(Y)
- CRF assigns a probability to entire sequences within the set of all possible output sequences $\mathcal{Y}$

$$\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} P(Y|X)$$

# Conditional Random Fields (CRF)

- CRF defines a function F which takes an input and output sequence and creates a feature vector (with K features)
- Probability of a tag sequence is then the log-linear sum of weighted features (similar to multinomial logistic regression)

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)$$

$$Z(X) = \sum_{Y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)$$

- Global feature vector $F_K$ is created by summing local feature $f_k$ at each index position in sequence Y (tags)

Pos

$$F_k(X,Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)$$

pos pos

- Linear chain CRF >> words at any position + tag $_i$ + tag $_{i-1}$

# Feature Sets

- Local feature values (for each sequence position i) are populated using a manually designed feature template
  - Feature templates are often coded as a python function
    get_local_features( sequence, position ) -> feature_set
  - Features can use information from be anywhere in the word sequence X (e.g. context window of 3 tokens either side of token)
  - Features sets can contain different value types (numeric, text, boolean)
  - Types of feature
    - Word
    - POS tag
    - Word shape type
    - Word prefix or suffix
    - Match to a lexicon (e.g. list of names) or gazetteer (e.g. list of cities)
- Feature values are summed over entire sentence X, which means there are always K features regardless of sentence length

# Feature Sets

- Example based on CRF NER lab python code

```
local_feature_dict = {
        'word' : word[i],
        'postag': postag[i],
        'word.lower()': word[i].lower(),
        'word.isupper()': word[i].isupper(),
        'word.istitle()': word[i].istitle(),
        'word.suffix': word[i].lower()[-3:],
        'word.islocation()': lookup_gaz(word[i]),
        '-1:word': word[i-1],
        '-1:postag': postag[i-1],
        '-1:word.lower()': word[i-1].lower(),
        '-1:word.isupper()': word[i-1].isupper(),
        '-1:word.istitle()': word[i-1].istitle(),
        '-1:word.isdigit()': word[i-1].isdigit(),
        '-1:word.suffix': word[i-1].lower()[-3:],
        '-1:postag[:2]': postag[i-1][:2],
        ...
}
```

**word**
**pos**

**word shape**

**suffix (3 char)**
**gazetteer lookup**
**previous word**

# Break

- Panopto Quiz - discussion point

- Which of these local feature types would be useful for a NER trying to label locations entities?

Word $X_i$
Word $X_{i-1}$
Word $X_{i+1}$
POS tag of $X_i$
POS tag of $X_{i-1}$
POS tag of $X_{i+1}$
Word $X_i$ capitalized?
Word $X_i$ all caps?
Word $X_i$ == 'London'
gazetteer lookup ($X_i$)

# Break

- Panopto Quiz - discussion point
- Which of these local feature types would be useful for a NER trying to label locations entities?

| | |
|---|---|
| Word $X_i$ | >> will overfit locations in training set |
| Word $X_{i-1}$ | |
| Word $X_{i+1}$ | |
| POS tag of $X_i$ | >> will capture POS noun phrase patterns |
| POS tag of $X_{i-1}$ | |
| POS tag of $X_{i+1}$ | |
| Word $X_i$ capitialized? | >> location names often Capitalized |
| Word $X_i$ all caps? | >> locations are not usually ALL CAPS YEAH? |
| Word $X_i$ == 'London' | >> will overfit to one location only |
| gazetteer lookup ($X_i$) | >> will allow match to any location in gazetteer list |

# Inference for NER

- Finding the best tag sequence for a given input X

$$\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} P(Y|X)$$

$$= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \frac{1}{Z(X)} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)$$

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)$$

from previous

# Inference for NER

- Finding the best tag sequence for a given input X

$$\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} P(Y|X) \qquad\qquad p(Y|X) = \frac{1}{Z(X)} \exp\left( \sum_{k=1}^{K} w_k F_k(X,Y) \right)$$

$$= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \cancel{\frac{1}{Z(X)}} \cancel{\exp} \left( \sum_{k=1}^{K} w_k F_k(X,Y) \right)$$

from previous

- We can ignore the exp() as we are interested only in relative ranking, and ignore the Z(X) as it will be constant for sequence X

from previous

$$F_k(X,Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i) \qquad\qquad \hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, X, i)$$

- Decode using Viterbi Algorithm (replace matrix A and B with global feature vector F)

# Evaluation of NER

- Community standard for evaluating taggers
  - POS tagging >> accuracy of tagged tokens
  - NER tagging >> micro P/R/F1 of entities (not tokens)
    >> macro P/R/F1 of entities (not tokens)

  - Reporting results per entity avoids bias to entities with more tokens
  - A choice is needed on how to treat partial entity matches (e.g. 'new york' for 'new york city')
  - Often the 'O' tag matches will be removed since the majority of any corpus will be 'O', and this will bias results reported using mean scores to the 'O' class performance

# Deep learning models for NER

- Modern deep learning NER approaches
  - Word representations - CBOW, skip-gram, word2vec, GloVE, BERT ...
  - Character representations - LSTM, GRU, CNN
  - Model - CNN, LSTM, GRU
  - Decoder - softmax, CRF
  - Performance - Ontonotes 0.92 F1 score >> compare to CRF NER lab !
- Academic NER
  - StanfordCoreNLP https://stanfordnlp.github.io/CoreNLP/
  - OSU Twitter NLP https://github.com/aritter/twitter_nlp
  - NeuroNER http://neuroner.com/
  - NERsuite http://nersuite.nlplab.org/
- Non-academic NER
  - spaCy https://spacy.io/api/entityrecognizer
  - NLTK https://www.nltk.org/book/ch07.html
  - OpenNLP https://opennlp.apache.org/
  - AllenNLP https://demo.allennlp.org/named-entity-recognition

# Required Reading

- Sequence Labelling for Parts of Speech
  - Jurafsky and Martin, Speech and Language Processing, 3rd edition (online) >> chapter 8

# Questions

- Panopto Quiz - 1 minute brainstorm for interactive questions

  Please write down in Panopto quiz in **1 minute** two or three questions that you would like to have answered at the next interactive session.

  Do it **right now** while its fresh.

  Take a screen shot of your questions and **bring them with you** at the interactive session so you have something to ask.