Lecture 24:
Information Extraction

Named Entity Types

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	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Figure 18.1 A list of generic named entity types with the kinds of entities they refer to.

These types were developed for the news domain as part of NIST's Automatic Content Extraction (ACE) program.

Other domains (e.g. biomedical text) require different types (proteins, genes, diseases, etc.)

Features for NER

Lists of common names exist for many entities

- Gazetteers (place names, www.geonames.org),
- Census-derived lists of first names and surnames,
- Genes, proteins, diseases, etc.
- Company names

Such lists can be helpful, but:

… **Zipf's Law:** these lists are typically not exhaustive, (and the distribution of names has a long tail)

… **Ambiguity:** many entity names either refer to different types of entities (*Washington*: person, places named after the person), or are used to refer to different types of entity (metonymy: *Washington* as reference to the US governement)

Feature-based NER

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Figure 18.5 Typical features for a feature-based NER system.

Train a sequence labeling model (MEMM or CRF), using features such as the ones listed above for English

- Word Shape: replace all upper-case letters with one symbol (e.g. “X”), all lower-case letters with another symbol (“x”), all digits with another symbol (“d”), and leave punctuation marks as is (“L’Occitane → “X’Xxxxxxxx”)
- Short Word Shape: remove adjacent letters that are identical in word shape “L’Occitane → “X’Xxxxxxxx” → “X’Xx”)

Neural NER

Sequence RNN (e.g. biLSTM or Transformer) with a CRF output layer.

Input: word embeddings, possibly concatenated with character embeddings and other features, e.g.:

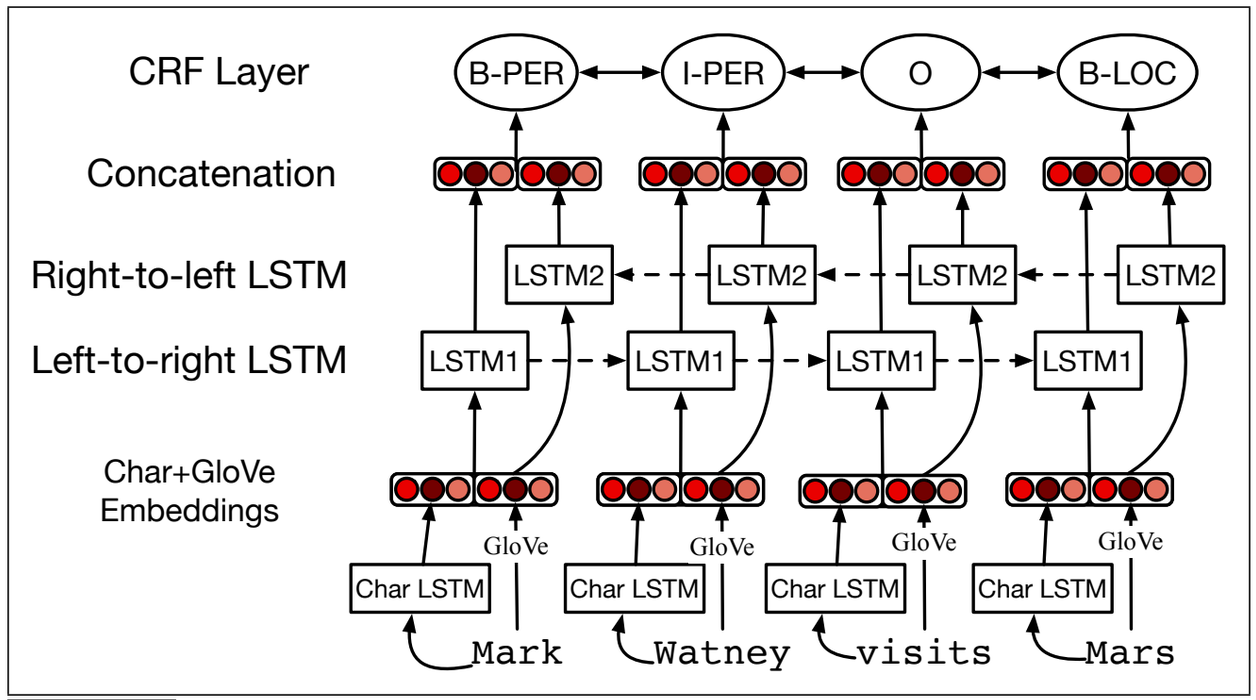

Figure 18.8 Putting it all together: character embeddings and words together in a bi-LSTM sequence model. After Lample et al. (2016).

Rule-based NER

The textbook gives an example of an iterative approach that makes multiple passes over the text:

- Pass 1: Use high-precision rules
  to label (a small number of) unambiguous mentions
- Pass 2: Propagate the labels of the previously detected
  named entities to any mentions
  that are substrings (or acronyms?) of these entities
- Pass 3: Use application-specific name lists
  to identify further likely names (as features?)
- Pass 4: Now use a sequence labeling approach for NER,
  keeping the already labeled entities
  as high-precision anchors.

The basic ideas behind this approach (label propagation, using high-precision items as anchors) can be useful for other tasks as well.