Named entity recognition (NER)

Problem definition

What sort of thing is "Bloopity-blod"? Person, location, organization, ...

- "I drove south from Bloopity-blod."
- "Bloopity-blod, the CEO of CapitalCorp, Inc."

NER applications

- NER for biomedical entities. i.e. are we talking about a disease, or a drug?

Is NER different from POS tagging?

Can't we just use more specific parts of speech,
e.g. replace NNP with PERSON, LOCATION, ORGANIZATION, ...?

"I drove south from Denver." **we need information about the surrounding parts of speech**

PRP VB RB IN __

"He grinned evilly at Barney."

Sentence structure is insufficient to allow good NER.

if we can somehow understand something about the implication of the surrounding words, then maybe we could do name entity recognition better

How do we do NER?

- word context
  - "I drove all the way from Bloopity-blod."
  - "Bloopity-blod, the CEO of CapitalCorp, Inc."
- word structure
  - "Mr. Blargyson" vs. "Blargingville" vs. "Blargitycorp"

maybe a person's name

maybe we haven't seen Blargitycorp before, but we saw piece of it before
—— need to break our words apart and look at sort of sequences of characters

some of the restrictive assumptions that we made were not horribly wrong for speechwriting, but are somewhat more detrimental for name recognition

Idea: POS tagging, but looser?

What if we clustered tokens according to similarity of the contexts in which they occur?

mutual information

> For n = 2 the maximum likelihood assignment of words to classes is equivalent to the assignment for which the average mutual information of adjacent classes is greatest.

That is, mutual information (roughly, correlation/predictability) is high when $P(c_i|c_{i-1})$ is high.

we don't even need a label (though we have one for Hutington), we just need to know that class5 and class67 tends to co-occur , and if we can find a good assignments of tokens to those classes, then we infer on this sort of mutual information principle that that's a good assignment of classes

Maximizing the average pairwise mutual information is intractible, so approximate...

E.g. treat(class 'Frizt') Huntington(class 'disease')
so the mutual information of this pair of adjacent classes is high because it's predictable
if we assigned Hutington to 'places' ——lower mutual information

so we tru assign tokens to classes in various ways, and eventually settle on the assignments that maximizez this pairwise mutual information

Brown clustering

1. assign a unique class to each token in the vocabulary
2. merge the two classes that will maximize the mutual information

mapping to the same tokens

- this can be done moderately efficiently

3. repeat 2

if we keep doing this forever, we'll end up with only one classes
all of the atoms in a vocabulary will be merge into one big class
and mutual information will be merged

Pro/Con

pro:

- we get an entire clustering tree and can choose what level(s) to work with
    - level 4 is very close to parts of speech

con:

- clusters are not easily interpretable
    - we will not get "person", "location", and "organization" clusters

Brown clusters and NER

- Need entity categories declarations
- Have Brown clusters which may be *close* to entity categories

-> supervised machine learning

ML for NER

Brown clusters are features (among others)

Use any old thing: SVM, Random Forest, etc.

BILOU

- **B**eginning
- **I**nside
- **L**ast
- **O**utside
- **U**nit

```
Mississippi and Philip Seymour Hoffman
U-LOC       O   B-PER  I-PER   L-PER
```