Machine Learning for Language Processing (L101)

Ann Copestake

Computer Laboratory
University of Cambridge

October 2017

Outline of today's lecture

NER overview

Maximum Entropy Models

NER in practice

Named Entity Recognition

- Identify all named entities in text

  Bill Gates says mosquitoes scare him more than sharks
  This reaction will produce 2,4-dinitrotoluene.
  This reaction will produce 2,4- and 2,6-dinitrotoluene.
- (usually) classify complete NE as PER, LOC etc
- NER is very important for many practical applications: search, information extraction, sentiment extraction ...
- Also as a preprocessor to parsing.

NER as an ML problem

```
Bill|I Gates|I says|O mosquitoes|O
scare|O him|O more|O than|O sharks|O
```

- Annotate tokens with I (in NER) or O (not in NER), or with a more complex scheme (e.g., IOB).
- Sequence classification (possibly multiple classifiers).
- Pretokenized input. POS tagging etc to supply features.
- Often highly complex set of features, including gazeteers, Wikipedia etc etc
- maybe hand-written rules (e.g., to help create training data)
- NER is VERY domain and genre dependent.

Simple IO:

`Bill|I Gates|I says|O mosquitoes|O scare|O him|O more|O than|O sharks|O`

IOB (also called BIO) with class labels:

`Bill|B-PER Gates|I-PER says|O mosquitoes|O scare|O him|O more|O than|O sharks|O`

- and others: BMEWO (beginning, middle, end, single word), BMEWO+ (adds tags to everything).
- The tagging scheme matters a lot for performance.
- Similar schemes in other contexts (e.g., character-based NN morphology models).
- The general case: nested NERs — essentially a form of parsing.

lingpipe-blog.com/2009/10/14/coding-chunkers-as-taggers-io-bio-bmewo-and-bmewo/

Maximum Entropy Model (MEM)

- MEM/MaxEnt is another name for multinomial logistic regression.
- MaxEnt is a discriminative classifier, especially useful when can't estimate full probabilities properly.
- Maximum Entropy Markov Models (MEMM): better for NER than HMM because allows for heterogeneous mix of features.
- Conditional Random Field (CRF) is an extension of MEMM.
- Slides in this section heavily based on J+M.

MEM schematically

$$P(c|\vec{f}) = \frac{1}{Z}\exp(\sum_i w_i f_i)$$

where Z normalizes, w_i is a weight and f_i is a numerically valued feature.

- actually w and f depend on class
- discriminative rather than generative

MEM vs NB

$$P(c|\vec{f}) = \frac{1}{Z}\exp(\sum_i w_i f_i) \qquad \text{(MaxEnt, schematic)}$$

$$P(c|\vec{f}) = \frac{\prod_{i=1}^{n} P(f_i|c)P(c)}{P(\vec{f})} \qquad \text{(NB)}$$

Linear regression: a recap

$$y = w_0 + \sum_{i=1}^{N} w_i \times f_i$$

Where w are weights and f are features.
Rewritten using an intercept feature, f_0, with value 1:

$$y = \sum_{i=0}^{N} w_i \times f_i$$

Weights chosen to minimize sum of squares of differences between prediction and observation.

Logistic regression: probabilistic classification

Abstractly we want (where f is the feature vector associated with observation x):

$$\begin{aligned} P(y = true|x) &= \sum_{i=0}^{N} w_i \times f_i \\ &= \vec{w} \cdot \vec{f} \end{aligned}$$

but what we're predicting won't be a probability.
Instead, we predict the log of the odds (logit function).

$$ln\left(\frac{P(y = true|x)}{1 - P(y = true|x)}\right) = \vec{w} \cdot \vec{f}$$

Logistic regression, continued

Classify observation as ‘true’ if:

$$P(y = true|x) > P(y = false|x)$$

That is:

$$\frac{P(y = true|x)}{1 - P(y = true|x)} > 1$$

or:

$$\vec{w} \cdot \vec{f} > 0$$

So logistic regression involves learning a hyperplane with true above and false below.

MaxEnt: Multinomial logistic regression

$$P(c|x) = \frac{1}{Z} \exp\left(\sum_{i=0}^{N} w_{ci} f_i\right)$$

where Z is the normalization factor:

$$Z = \sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i\right)$$

MaxEnt: Multinomial logistic regression

with numerical-valued features:

$$P(c|x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i\right)}$$

MaxEnt: Multinomial logistic regression

with booean-valued features:

$$P(c|x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i(c, x)\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i(c', x)\right)}$$

Features include the class:

$$\begin{aligned} f_1(c, x) &= 1 \text{ if } \textit{word}_i \text{ ends in “ic” \& } c = \text{CJ} \\ &= 0 \text{ otherwise} \end{aligned}$$

Training and using MaxEnt models

- MaxEnt can be used for hard classification: in effect, a linear expression that separates class from other classes.
- but MaxEnt also gives a probability distribution: necessary for sequence classification.
- Training maximizes the log likelihood of the training samples (but regularization to penalize large weights).
- Training process makes no assumptions beyond data: model should fit constraints and have maximum entropy.
- Equivalent to maximizing the likelihood for multinomial logistic regression.

MaxEnt Markov Model: MEMM

- Viterbi (as HMM) for most probable sequence of classes.
- MEMM vs HMM (assuming bigram features).

$$P(Q|O) = \prod_{i=1}^{n} P(q_i|q_{i-1}, o_i) \qquad \text{(MEMM)}$$

$$P(Q|O) = \prod_{i=1}^{n} P(o_i|q_i) \times \prod_{i=1}^{n} P(q_i|q_{i-1}) \qquad \text{(HMM)}$$

where Q is state sequence and O is observations.

- But MEMM can use much more complex features.

NER: state of the art

- CRF (Conditional Random Field), introduced in 2001. Global normalization of probabilities: theoretically better than MEMM (practically not always much difference, slower to train).
- Recently, various LSTM models proposed: much cleaner, less domain-dependent, don't need external gazeteers, performance at least as good as best previous models.
- Small, limited standard test sets, still quite low performance for some languages.

Annotating NERs

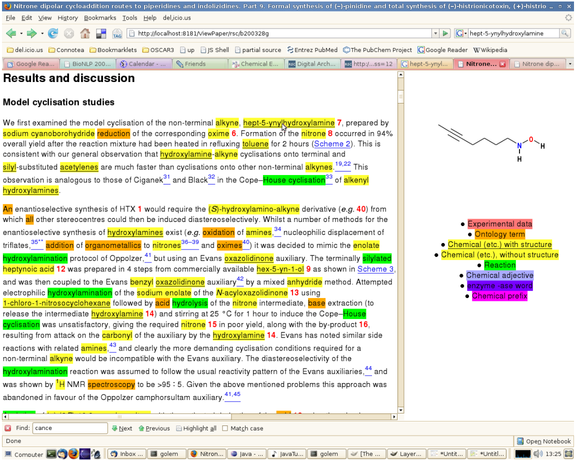
- Deciding on span:

  The New York Stock Exchange fell today.

  New York Stock Exchange or The New York Stock Exchange?
- Nested or overlapping NEs?

  The New York Stock Exchange fell today.
  The New York and Chicago Stock Exchanges fell today.
- Named entity or ordinary noun phrase?
  Queen Elizabeth, *the Queen*, *the Queen of England*, *the queen of England*, *a queen of England*.

Chemistry NERs (Corbett, Murray-Rust et al)



Nitrone dipolar cycloaddition routes to piperidines and indolizidines. Part 9. Formal synthesis of (–)-pinidine and total synthesis of (–)-histrionicotoxin, (+)-histrio

Results and discussion

Model cyclisation studies

We first examined the model cyclisation of the non-terminal alkyne, hept-5-ynylhydroxylamine 7, prepared by sodium cyanoborohydride reduction of the corresponding oxime 6. Formation of the nitrone 8 occurred in 94% overall yield after the reaction mixture had been heated in refluxing toluene for 2 hours (Scheme 2). This is consistent with our general observation that hydroxylamine-alkyne cyclisations onto terminal and silyl-substituted acetylenes are much faster than cyclisations onto other non-terminal alkynes.[19,22] This observation is analogous to those of Ciganek[31] and Black[32] in the Cope–House cyclisation[33] of alkenyl hydroxylamines.

An enantioselective synthesis of HTX 1 would require the (*S*)-hydroxylamino-alkyne derivative (*e.g.* 40) from which all other stereocentres could then be induced diastereoselectively. Whilst a number of methods for the enantioselective synthesis of hydroxylamines exist (*e.g.* oxidation of amines,[34] nucleophilic displacement of triflates,[35**] addition of organometallics to nitrones[36–39] and oximes[40]) it was decided to mimic the enolate hydroxylamination protocol of Oppolzer,[41] but using an Evans oxazolidinone auxiliary. The terminally silylated heptynoic acid 12 was prepared in 4 steps from commercially available hex-5-yn-1-ol 9 as shown in Scheme 3, and was then coupled to the Evans benzyl oxazolidinone auxiliary[42] by a mixed anhydride method. Attempted electrophilic hydroxylamination of the sodium enolate of the *N*-acyloxazolidinone 13 using 1-chloro-1-nitrosocyclohexane followed by acid hydrolysis of the nitrone intermediate, base extraction (to release the intermediate hydroxylamine 14) and stirring at 25 °C for 1 hour to induce the Cope–House cyclisation was unsatisfactory, giving the required nitrone 15 in poor yield, along with the by-product 16, resulting from attack on the carbonyl of the auxiliary by the hydroxylamine 14. Evans has noted similar side reactions with related amines,[43] and clearly the more demanding cyclisation conditions required for a non-terminal alkyne would be incompatible with the Evans auxiliary. The diastereoselectivity of the hydroxylamination reaction was assumed to follow the usual reactivity pattern of the Evans auxiliaries,[44] and was shown by ^{1}H NMR spectroscopy to be >95 : 5. Given the above mentioned problems this approach was abandoned in favour of the Oppolzer camphorsultam auxiliary.[41,45]

Chemistry NER (Corbett and Copestake, 2008)

- Used cascaded classifiers: preclassifier (character ngrams), first-order MEMM, entity type rescorer.
- Complex feature examples:

  `4G=ceti`

  the character sequence ‘c’ ‘e’ ‘t’ ‘i’ is in the token

  `bg:0:1:ct=CJ_w=acid`

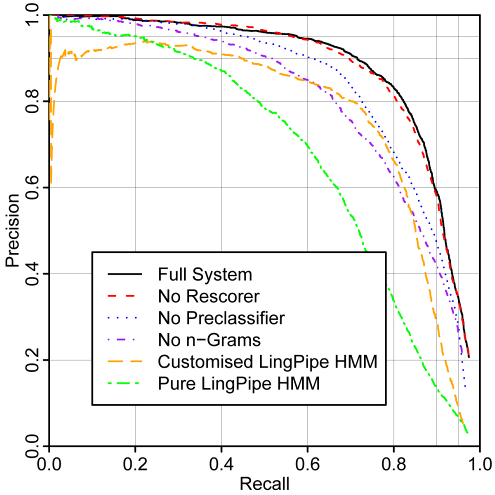
  token is of type CJ (chemical adjective) according to preclassifier and next token is ‘acid’
- Use probability estimates to experiment with precision vs recall.

Precision and recall

- Precision: percentage of NERs found that were correct
- Recall: percentage of annotated NERs that were found
- F-measure: combined precision and recall

$$F_1 = \frac{2PR}{P + R}$$

Chemistry NERs: precision and recall



Beyond F_1

Confidence scores allow precision/recall to be varied:

- High precision: good where high redundancy but high cost to checking result. e.g., normal search
- High recall: good where little or no redundancy, false positives not as important as false negatives.
  e.g., exhaustive search
  e.g., chemistry NER as preprocessor to parsing — because unrecognised NER leads to very bad parse results

Next time

- Your next session is Tuesday 17th at 12, seminar with Ted.
- My next lecture is Thursday 19th at 3pm (kernels and perceptrons).