Learning maximum-entropy models of salience via EM

Micha Elsner
joint work with Eugene Charniak and Mark Johnson

Department of Computer Science
Brown University

September 30, 2009

Motivation

The White Queen looked timidly at Alice, who felt she ought to say something kind, but really couldn't think of anything at the moment.

- Pronouns are potentially ambiguous.
- Does she mean Alice, or the White Queen?
- Technically could be either, but strong intuitions.

Starting point: machine translation

IBM model 2

Generate German from English:

- *Align:* pick a random English word to translate.
- *Translate:* pick an appropriate German word.

English:	He	can	sing	well
German:	Er	kann	gut	singen

Our generative setting

- "Translate" the context into a pronoun...
  - Via a hidden alignment.

Source text: The White Queen looked at Alice who felt

Target text: she

The "translation" model (Charniak+Elsner '09)

Pronouns uniquely identified by:

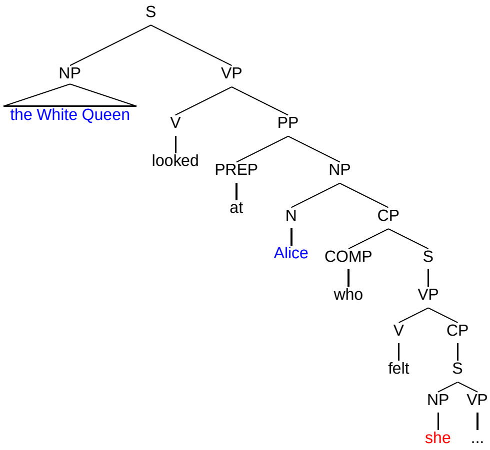
- Person (I/you/it)
- Number (it/they)
- Gender of singular pronouns (he/she/it)
  - English plural pronouns ("they") unmarked for gender.

$P(pro|ante)$ modeled as:

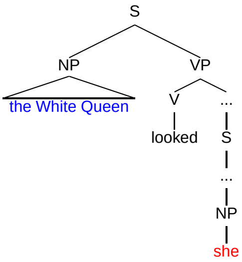
$$P(pers(pro)|pers(ante)) \times$$
$$P(num(pro)|num(ante)) \times$$
$$\sum_{possible\ gen(pro)} P(gen(pro)|gen(ante))$$

Modeling alignment: issues

The White Queen and Alice: both feminine singular, so translation model doesn't help us.
Need alignment function based on the syntax.



Features



- syntactic role: subject
- position: beginning of sentence
- proximity: same sentence
- within-sentence proximity: 6 words away
- phrase type: proper noun phrase
- determiner: “the”
- head word: “Queen”

The alignment function

Each pronoun i has set of possible antecedents A_i.
A noun phrase a has some features $S(a, i)$.

Alignment function:

$$P(ante(i) = a \in A_i \mid S(a, i), \{S(A_i, i)\})$$

The ugly method (Charniak+Elsner '09)

$$\begin{aligned} P(ante(i) = a \mid S(a,i), \{S(A_i)\}) = \\ P(ante(i) = a \mid S(a,i)) \sim \\ Bernoulli(\bullet; \theta_{S(a,i)}) \end{aligned}$$

For every possible antecedent, flip a coin to decide if it's the true antecedent. Just **assume** one, and only one, coin will come up heads.

The ugly method (Charniak+Elsner '09)

$$\begin{aligned} P(ante(i) = a \mid S(a,i), \{S(A_i)\}) = \\ P(ante(i) = a \mid S(a,i)) \sim \\ Bernoulli(\bullet; \theta_{S(a,i)}) \end{aligned}$$

For every possible antecedent, flip a coin to decide if it's the true antecedent. Just **assume** one, and only one, coin will come up heads.

- Not probabilistically legitimate
- One parameter θ for each possible feature vector $S(a,i)$: can't be too sparse

Using log-linear models

A more standard approach:

$$P(ante(i) = a \mid S(a,i), \{S(A_i)\}) = \frac{exp(w \bullet S(a,i))}{Z}$$
$$Z = \sum_{x \in A_i} exp(w \bullet S(x,i))$$

Like softmax multilabel classification, but the set of 'labels' is different for every datapoint.

Using EM

Log-linear form specifies a *conditional* distribution...
Part of overall *generative* model.

Simple EM algorithm

- E-step: compute probabilities $P(ante(i) = a)$

  and sum to compute $E\left[S(ante(i), i), \{S(A_i, i)\}\right]$

  ...the expected number of times we pick an antecedent with features S from a set of available phrases with features $\{S\}$
- M-step: estimate w by gradient descent on the likelihood

Faster inference?

Problem: there are a *lot* of sufficient statistics:

$$E\left[S(ante(i), i), \{S(A_i, i)\}\right]$$

...and feature vector S is probably sparse.

Possibility: online perceptron-style updates:
Stepwise EM ((Sato+Ishii '00) and (Liang+Klein '09)):

- Compute expectations for a batch of examples
- Estimate the gradient w' and update $w = \eta w + (1 - \eta) w'$

Getting the batch size and learning rate right is tricky...

Preliminary results

Initialized the max-ent alignment to the distribution learned by the previous system.

system	performance	# of alignment params
(Charniak+Elsner '09)	67.2	2592
my reimplementation	65.4	2592
max-ent	65.7	61

- There is a compact representation of the alignment function
- It occurs near a local max of the (legitimate) likelihood

Why no improvement?

- Max-ent alignment could be similar to the “ugly” distribution...
  - if partition function Z for each example approximately equal

Would imply:

Most syntactic environments have approximately same amount of important noun phrases.

Haven’t tested this!

Same-head coreference

Most NPs with the same *head word* are coreferent:

Alice thought to herself... Alice said...

But some are not:

the White Queen ... the Red Queen...

one day at a time ... the day before...

it sighed and the consequence was...
it wouldn't come out and the consequence was...

Modeling idea

Generate the NPs from left to right...

Alignment

- Max-ent produces coreferent NPs
- Uniform distribution produces others

$$P(ante(i) = a \mid S(a,i), \{S(A_i)\}) \propto \lambda * exp(w \bullet S(a,i)) + (1-\lambda) * \frac{1}{|S|}$$

Translation model

Input: antecedent NP
Output: similar NP with different modifiers

Really, really preliminary results

Pronoun model plus model for NPs with same heads:

	link all	our model
cluster overlap	69	**74**
link precision	54	**65**
link recall	**50**	35
f-score	**52**	45

Better cluster overlap, but trades recall for precision.

Future directions

Current goals:

- Better tuning for perceptron-style updates
- Analysis of different roles of translation/alignment
- Link NPs with different heads

Thanks for listening!

Please ask questions, or contact me:

melsner@cs.brown.edu