A Maximum Entropy Based Honorificity Identification for Bengali Pronominal Anaphora Resolution

Apurbalal Senapati and Utpal Garain

Presented by Samik Some

Introduction

- Pronomial anaphora resolution is important for several NLP tasks such as question answering, summarization etc.
- In English pronouns carry gender information such as he/she/it which is helpful when resolving them to find the nouns they refer to.
- In Bengali pronouns do not have gender information however there is honorificity information embedded in pronouns. eg. Three forms of the verb খাওয়া; খা, খাও and খান.
- Such honorific information is present in both written and spoken forms, and may be used for resolving pronoun references.
- There are multiple features which contribute to determining noun honorificity in Bengali which led to the selection of a maximum entropy model.
- Such a model offers better performance than earlier methods.

Honorific Information in Bengali

- Three types of honorificity exist in Bengali language both in written and spoken form.
- The highest degree (SUP class) refers to people who are of higher status or generally respectable people such as doctors, teachers, elders, etc.
- Next is the neutral form (NEU class) which is used when referring to close family members, children, younger family members, etc.
- The lowest level (INF class) is used for very close friends, very close relations, or people who are presumed to be of lower social status such as rickshaw pullers, housemaids, etc.
- Such honorific information can be extracted from several sources such as the title placed before or after a name.
- It can also be obtained by observing the inflection of the main verb.
- Pronouns also conform with such honorificity and there are different forms for different levels.

Maximum Entropy Modelling

A training sample is summarized in terms of its empirical probability distribution. Here x is the context and y is the corresponding class.

$$\tilde{p}(x, y) = \frac{1}{N} \times \text{number of times (x, y) occurs in the sample}$$

We define events in terms of binary values feature functions which gives 1 if the context x contains some particular information and the corresponding class is y.

$$f(x, y) = 1/0 \text{ depending on context and class}$$

The statistic we are interested in is the expected value of the feature function with respect to the empirical distribution p~(x,y).

$$\tilde{p}(f) \equiv \sum_{x,y} \tilde{p}(x, y) f(x, y)$$

Maximum Entropy Modelling (contd.)

The expected value that the model p(y|x) assigns to f is given as...

$$p(f) \equiv \sum_{x,y} \tilde{p}(x)p(y|x)f(x,y)$$

We require that our model conforms to the following restriction.

$$p(f) = \tilde{p}(f)$$

Thus the solution space for our model is...

$$C = \{p \in P | p(f\mathrm{i}) = \tilde{p}(f_i)\ for\ i \in \{1,2,\dots,n\}\}$$

We can find the most uniform model in this space by maximizing the following entropy function.

$$H(p) \equiv -\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x)$$

Feature functions

- A 10-dimensional feature vector is used.
- $f_1(x, y) = 1$ when y is SUP and x is the person and honorific addressing term is in the left position; 0 otherwise.
- $f_2(x, y) = 1$ when y is SUP and x is the person and honorific addressing term is in right position; 0 otherwise.
- $f_3(x, y) = 1$ when y is SUP and x is the person and honorific addressing terms are both in left and right positions; 0 otherwise.
- $f_4(x, y) = 1$ when y is SUP and x is the person and honorific addressing term is in left position and the main verb associated with x is in SUP class; 0 otherwise.
- $f_5(x, y) = 1$ when y is SUP and x is the person and honorific addressing term is in right position and the main verb associated with x is in SUP class; 0 otherwise.

Feature functions (contd.)

- f6(x, y) = 1 when y is SUP and x is the person and honorific addressing terms are both in left and right positions and the main verb associated with x is in SUP class; 0 otherwise.
- f7(x, y) = 1 when y is SUP and x is the person and the main verb associated with x is in SUP class; 0 otherwise.
- f8(x, y) = 1 when y is NEU and x is the person and the main verb associated with x is in NEU class; 0 otherwise.
- f9(x, y) = 1 when y is NEU and x is the person and there is neither left honorific terms nor right honorific terms and the main verb is absent; 0 otherwise.
- f10(x, y) = 1 when y is INF and x is the person and no honorific term is either in left or in right position and the main verb is in INF class; 0 other-wise.

Training

- The feature functions are computed locally within a sentence and the model was trained using a large Bengali corpus (35 million words)
- The data was defined in the form (x,y) where x is the sentence or phrase containing a person with context information and y is the honorific information.

6	মিস		XC	B-NP	B-PERSON	-
7	আগাথা		XC	I-NP	I-PERSON	SUP
8	হ্যারিসন		NNP	I-NP	I-PERSON	SUP
9	প্রথমে		NN	B-NP	o	-
10	শান্তিনিকেতনে	NNP	B-NP	o	-	
11	আসেন		VM	B-VGF	o	SUP
12	1930		XC	B-NP	o	-
13	সালে		NNP	I-NP	o	-
14	।		SYM	I-NP	o	-

Format of training data

Data description	Number
#text	25
#words	48,177
#persons	1,661
#SUP category	1,227
#NEU category	288
#INF category	146

Coverage

Results

Tested on an extended version of ICON 2011 annotated data for anaphora resolution in Bengali. The results are as shown.

Data description	Number
#text	13
#words	27,454
#persons	1,243
#SUP category	901
#NEU category	236
#INF category	106

Coverage

Category	SUP	NEU	INF	Total
SUP	810	91	0	901
NEU	29	207	0	236
INF	5	14	87	106
Total	844	312	87	1243

Category	P	R	F1
SUP	95.97	89.90	92.83
NEU	66.34	87.71	75.54
INF	100.00	82.07	90.15

Honorificity detection

Metric		Baseline	System I	System II	System III
MUC	P	0.437	0.477	0.489	0.538
	R	0.426	0.426	0.462	0.605
	F1	0.431	0.450	0.475	0.569
B^3	P	0.577	0.667	0.678	0.740
	R	0.676	0.786	0.832	0.842
	F1	0.608	0.721	0.747	0.787
CEAFM	P	0.614	0.614	0.654	0.786
	R	0.602	0.602	0.646	0.695
	F1	0.608	0.607	0.650	0.737
CEAFE	P	0.661	0.771	0.797	0.804
	R	0.601	0.601	0.642	0.555
	F1	0.630	0.675	0.712	0.656
BLANC	P	0.480	0.500	0.542	0.765
	R	0.582	0.628	0.678	0.771
	F1	0.526	0.556	0.603	0.767
Avg.	F1	0.560	0.602	0.637	0.703

Pronomial anaphora resolution

Thank you