Institute for Natural Language Processing

Data-driven Multilingual Coreference Resolution using Resolver Stacking

Available at
http://www.ims.uni-stuttgart.de/~anders/

Anders Björkelund and Richárd Farkas
Institute for Natural Language Processing, University of Stuttgart

Approach

- Mention detection
  - Non-referential classifier
- Coreference classifier
  - Heavy feature engineering
  - Disallowing transitive nesting
  - Cluster mention decorder
  - Resolver stacking

Mention Detection

Arabic: NP + PRP + PRP$
Chinese: NP + PN + NR
English: NP + PRP + PRP$ + NEs - NonRef

	$th = 0.5$			$th = 0.95$			
	Precision	Recall	F_1	Precision	Recall	F_1	# occurrences
it	75.41	61.92	68	86.78	38.65	53.48	10,307
we	65.93	41.61	51.02	75.41	24.20	36.64	5,323
you	79.10	74.26	76.60	88.36	51.59	65.15	11,297
Average	75.73	63.05	68.81	86.17	41.04	55.60	26,927

Table 1: Performance of the non-referential classifier used for English. Precision, recall, and F-measure are broken down by pronoun (top three rows), and the micro-average over all three (bottom row). The left side uses a probability threshold of 0.5, and the right one a threshold of 0.95. The last column denotes the number of occurrences of the corresponding token. All numbers are computed on the development set.

Decoders and Stacking

- BestFirst (BF),
- Pronouns Closest First (PCF),
- Cluster mention decoder (AMP):

$$score(m_i, m_j) = \Big(\prod_{m_c \in C} P(coref|(m_c, m_j))\Big)^{1/|C|}$$

- Stacking: AMP + (BF/PCF)

Arabic	BF	PCF	AMP	Stacked
MD	58.63	58.49	58.21	60.51
MUC	45.8	45.4	43.2	46.66
BCUB	66.65	66.56	66.39	66.3
CEAFE	41.52	41.58	43.1	42.57
CoNLL	51.32	51.18	50.9	51.84
Chinese	BF	PCF	AMP	Stacked
MD	67.22	67.19	66.79	67.61
MUC	59.58	59.43	57.23	59.84
BCUB	72.9	72.82	72.7	73.35
CEAFE	46.99	46.98	48.25	47.7
CoNLL	59.82	59.74	59.39	60.30
English	BF	PCF	AMP	Stacked
MD	74.33	74.42	73.75	74.96
MUC	66.76	66.93	62.74	67.12
BCUB	70.96	71.11	68.05	71.18
CEAFE	45.46	45.83	46.49	46.84
CoNLL	61.06	61.29	59.09	61.71

Transitive Nesting

(1) ... she seemed to have such a good relationship with [[her]$_b$ mother]$_a$. Like [[her]$_d$ mother]$_c$ treated her like a human being ...

(2) [[Taiwan]$_f$'s]$_e$

Modified decoder to disallow transitive nesting, e.g. Skip linking (*a*,*d*), if (*c*,*d*) was negative

Official Results

2nd place in Shared Task!

Arabic	PM	GB	GM
MD	60.55	60.61	76.43
MUC	47.82	47.90	60.81
B^3	68.54	68.61	67.29
CEAFE	44.3	44	49.32
CoNLL	53.55	53.50	59.14
Chinese	PM	GB	GM
MD	66.37	71.02	83.47
MUC	58.61	63.56	76.85
B^3	73.10	74.52	76.30
CEAFE	48.19	50.20	56.61
CoNLL	59.97	62.76	69.92
English	PM	GB	GM
MD	75.38	75.3	86.16
MUC	67.58	67.29	78.70
B^3	70.26	69.70	72.67
CEAFE	45.87	45.27	53.23
CoNLL	61.24	60.75	68.20

Table 3: Performance on the shared task test set. Using predicted mentions (PM; i.e., the official evaluation), gold mentions boundaries (GB), and gold mentions (GM).

Feature Set

	Arabic	**Chinese**	**English**
Alias			•
J_{-1}POS		•	
JDemonstrative			•
IBOLemma	•		
IParCat		•	•
IParSubCat	•	•	•
ISubCat		•	
IHdLC	•		
IHdLemma	•		
IHdPos	•	•	
IHd_{-2}Lemma	•		
INE			•
I_{+1}POS			•
I_{-1}Form		•	•
I_{-1}POS	•		
IForm	•	•	•
DSPathHdForm			•
DSPath		•	
SSPath	•	•	
SSPathHdPos	•		•
StringMatch			•
SentDist		•	
Nested		•	
IJBWUVEditScript-10	•		
IJFormEditScript-10	•		
IJFormEditDistance	•		
IJBWUVEditScript+IParSubCat-10	•		
JBOBWUV+IHdBWUV	•		
JSubCat+Nested		•	
JHdLemma+IHdPos	•		
JHdPos+IHdLemma	•		
J_{first}Form+IHdForm		•	•
J_{first}Pos+I_{+1}Form			•
JForm+IForm		•	•
IBOLemma+JHdLemma	•		
IParSubCat+JHdForm		•	
IParSubCat+$JHdForm_{prp}$			•
IParSubCat+J_{-1}Pos		•	
ISubCat+Nested			•
IGender+$JHdForm_{prp}$			•
IHdForm+JHdForm		•	•
IHd_{+1}Form+IHdPos			•
IHdPos+$JHdForm_{prp}$			•
IHdPos+IHd_{-1}Pos	•		
$IHdForm_{prp}$+$JHdForm_{prp}$	•		
I_{+1}Pos+$JHdForm_{prp}$			•
I_{-1}Form+JHdForm			•
I_{-1}Pos+JHdLemma	•		
I_{-1}Pos+IParSubCat	•		
SSPath+$JHdForm_{prp}$		•	•
SSPath+Genre		•	
StringMatch+IProperName	•		•
SentDist+JHdForm		•	
SentDist+JPronoun	•	•	•
SentDist+$JHdForm_{prp}$			•
StringMatch+JPronoun		•	
StringMatch+Distance		•	
Genre+IHdForm		•	
Genre+I_{first}Form		•	
Genre+Nested		•	•
MentDistance+JPronoun	•		•
Nested+JPronoun			•
SameSpeaker+$IHdForm_{prp}$+$JHdForm_{prp}$		•	•
JQuoted+$JHdForm_{prp}$+IDominatingVerb	•		
IParSubCat+MentDistance+JPronoun			•
SSPath+JPronoun+IPronoun	•		
Genre+$IHdForm_{prp}$+$JHdForm_{prp}$			•
MentDistance+JPronoun+IParSubCat-10	•		
StringMatch+IProperName+IHdForm+JHdForm			•
JHdSSMatch+JProperName+IProperName+MentDistance+IPronoun		•	
Genre+Nested+JHdSSMatch+JProperName+IProperName		•	

Legend:
I, **J** – Mention I or J
POS, **Form**, **Lemma**, **BWUV**, **LC** – Part-of-speech tag, surface form, Lemma, Buckwalter unvocalised form, Last Character of surface form
SubCat, **ParSubCat** – Subcategorzation frame in the syntax tree, SubCat of parent node
$_{\pm 1/2}$ – Applied tokens outside the span, one or two tokens before or after
$_{prp}$ – Only fires for surface forms if they are pronouns
$_{first}$ – The first token in a span
-10 (suffix) – Means only features that occur more than 10 times are included
SSPath, **DSPath** – Path in syntax tree when I and J occur in same sentence (SS), or in different sentences (DS)
SSMatch – Substring match
BO – Bag of ...
NE – Named Entity

Additional Experiments

- Training on train+dev only minor improvement (Chinese, English)
- Training on gold syntax and testing on predicted is harmful (Arabic, Chinese, English)
- When testing on gold syntax, the models trained on predicted syntax are much better (Chinese, Enligsh)
- Gold boundaries are worse than predicted boundaries, even with gold syntax in test data (English)
- Ask for handout with detailed tables!