Automatic Classification of Communicative Functions of Definiteness

Carnegie Mellon University

Archna Bhatia[LTI], Chu-Cheng Lin[LTI], Nathan Schneider[LTI], Yulia Tsvetkov[LTI], Fatima Talib Al-Raisi[LTI], Laleh Roostapour[LTI], Jordan Bender[P], Abhimanu Kumar[LTI], Lori Levin[LTI], Mandy Simons[PH], and Chris Dyer[LTI]

Language **T**echnologies **I**nstitute / Department of **Ph**ilosophy, Carnegie Mellon University • Department of Linguistics, University of **P**ittsburgh. Pittsburgh, PA

Abstract

Definiteness expresses a constellation of semantic, pragmatic, and discourse properties (the *communicative functions*) of an NP. Our supervised classifier for English NPs uses lexical, morphological, and syntactic features to predict the communicative functions in terms of a language-universal classification scheme and establishes strong baselines for future work. Additionally, analysis of the features and learned parameters in the model provides insight into the grammaticalization of definiteness in English, not all of which is obvious a priori.

Communicative Functions of Definiteness (CFD) Annotation Scheme

Nonanaphora [−A,−B]	**999**
Unique [+U]	**287**
∗ **Unique_Hearer_Old** [+F,−G,+S]	**251**
Unique_Physical_Copresence [+R]	13
Unique_Larger_Situation [+R]	237
Unique_Predicative_Identity [+P]	1
∗ Unique_Hearer_New [−F]	36
Nonunique [−U]	**581**
∗ **Nonunique_Hearer_Old** [+F]	**169**
Nonunique_Physical_Copresence [−G,+R,+S]	39
Nonunique_Larger_Situation [−G,+R,+S]	117
Nonunique_Predicative_Identity [+P]	13
∗ Nonunique_Hearer_New_Spec [−F,−G,+R,+S]	231
∗ Nonunique_Nonspec [−G,−S]	181
Generic [+G,−R]	**131**
∗ Generic_Kind_Level	0
∗ Generic_Individual_Level	131

Anaphora [+A]	**1574**
Basic_Anaphora [−B,+F]	**795**
∗ Same_Head	556
∗ Different_Head	329
Extended_Anaphora [+B]	**779**
∗ Bridging_Nominal [−G,+R,+S]	43
∗ Bridging_Event [+R,+S]	10
∗ Bridging_Restrictive_Modifier [−G,+S]	614
∗ Bridging_Subtype_Instance [−G]	0
∗ Bridging_Other_Context [+F]	112
Miscellaneous [−R]	**732**
Pleonastic [−B,−P]	53
Quantified	248
Predicative_Equative_Role [−B,+P]	58
Part_Of_Noncompositional_MWE	100
Measure_Nonreferential	125
Other_Nonreferential	148

Examples for Communicative Functions

CFD Label	Example
Unique_Physical_Copresence	**John** here is an investment banker.
Unique_Larger_Situation	In the days since **Hillary Clinton** unburdened herself in an interview with The Atlantic's Jeffrey Goldberg ...
Unique_Predicative_Identity	Clark Kent is **Superman**.
Unique_Hearer_New	a restaurant chain named **Shoney's**
Nonunique_Physical_Copresence	**The podium** is too high.
Nonunique_Larger_Situation	**the chair** (at a conference) / **today**
Nonunique_Predicative_Identity	He is **the manager**.
Nonunique_Hearer_New_Specific	I am looking for **a nurse**. Her name is Sara.
Nonunique_Nonspec	I am looking for **a nurse** [any nurse would do].
Generic_Kind_Level	**Dinosaurs** are extinct.
Generic_Individual_Level	**Cats** have fur.
Basic_Same_Head	I'm going to tell you a quick story. It's **a true story**.
Basic_Different_Head	I adopted a cat this weekend. **The animal** is so cute.
Extended_Bridging_Nominal	I looked at an apartment yesterday. **The kitchen** was really large.
Extended_Bridging_Event	My friend's son got married this weekend. **The bride** looked beautiful.
Extended_Bridging_Restrictive_Modifier	**the house** next door/ John's **daughter**
Extended_Subtype_Instance	I collect coins. I have **a 1943 steel penny**.
Extended_Other_Context	I want to focus on what many of you have said you would like me to elaborate on. What can you do about **the climate crisis**?
Pleonastic	**It is raining.**
Quantified	**All the people** / **no motorcade**
Predicative_Equative_Role	He's **a teacher**. / This is **an opportunity**.
Part_of_Noncompositional_MWE	Ole' Charlie kicked **the bucket** today.
Measure_Nonreferential	**hours** later / **miles** away
Other_Nonreferential	**global warming** / **concern** / the topic of **energy**

Analysis Example: Specificity

Confirmation of known facts	Hypotheses to test
+ the definite article "the" + possessives (PRP$) + proper nouns (NNP) + 2nd person pronouns + NPs with "the" as the first dependent	+ objects of "from" + NPs with NNP as their last dependent + NPs with possessive pronouns immediately preceding the head (rather than the ones with intervening words)
− the indefinite article "a"	− NPs with comparative adjectives (JJR)

Acknowledgements

This research was sponsored by a grant from the U.S. Army Research Lab and the U.S. Army Research Office.

Classification Model

We use an in-house implementation of a multiclass logistic regression classifier.

Feature function: $\mathbf{f}(x,y) = \underbrace{\boldsymbol{\phi}(x)}_{\substack{\text{percepts}\\ \text{(input features)}}} \times \underbrace{\tilde{\boldsymbol{\omega}}(y)}_{\substack{\text{attributes}\\ \text{(output features)}}}$

Learning objective (L_2-regularized):

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} -\lambda||\boldsymbol{\theta}||_2^2 + \sum_{\langle x,y\rangle \in \mathscr{D}} \log \frac{\exp \boldsymbol{\theta}^\top \mathbf{f}(x,y)}{\sum_{y' \in \mathscr{Y}} \exp \boldsymbol{\theta}^\top \mathbf{f}(x,y')}$$

Features

Words of Interest Head of the NP, its dependents, its governor (external to NP), its first ancestor verb — token, lemma, POS tag, dependency relation, a binary indicator of plurality on the head N, first_dependent, last_dependent, auxiliaries of the first ancestral verb, first ancestral verb with a negative particle as dependent.

Structural — path length to the root, path length to the first ancestral verb, number of dependents, number of dependency relations that link non-neighbors.

Positional — token length of the NP, NP's location in the sentence (first or second half), the first ancestral verb's position relative to the head (left or right), POS & lemma of the left and the right neighbors of the head, governor, and the first ancestral verb.

Above features of NPs in Following NP-NP relation Types immediate parent, immediate child, immediate precedent, immediate successor, the nearest preceding coreferent mention.

Accuracy

Condition	Params #	Exact %	Soft %
Majority baseline	—	12.1	47.8
Log-linear			
attributes	473,064	38.7	77.1
labels	413,931	40.8	73.6
attributes, labels	926,417	43.7	**78.2**
Random forest	20,363	**49.7**	77.5

Communicative Function Label Accuracy

CFD label	#	F_1 (%)
Bridging_Restrictive_Modifier	552	68
Same_Head	452	41
Different_Head	271	32
Quantified	213	57
Nonunique_Hearer_New_Specific	190	40
Nonunique_Nonspec	173	13
Other_Nonreferential	134	37
Generic_Individual_Level	113	13
Measure_Nonreferential	98	40
Unique_Larger_Situation	97	55
Nonunique_Larger_Situation	97	27
Bridging_Other_Context	96	11
Part_of_Noncompositional_MWE	88	18
Predicative_Nonidentity	57	—
Pleonastic	44	88
Nonunique_Physical_Copresence	36	—
Bridging_Nominal	33	15
Unique_Hearer_New	26	—
Nonunique_Predicative_Identity	10	—
Bridging_Event	9	—