Title, author list not final. Page limit: 8 + 2 bib (will be granted more space in CR)

A Classifier for Communicative Functions of Definiteness

Abstract

Whether a NP is realized grammatically as definite or not depends on a variety of semantic, pragmatic, and discourse criteria, or *communicative functions*, the interaction of which varies from language to language. We present a supervised classifier for English that uses lexical, morphological, and syntactic features to predict communicative functions of definiteness. The benefits of this work are twofold: linguistically, the classifier's features and weights model the *grammaticalization* of definiteness in English, not all of which are obvious. Computationally, it presents a framework to predict *semantic and pragmatic* communicative functions of definiteness which, unlike lexical and morphosyntactic features, are preserved in translation. The classifier may therefore be useful for MT and other NLP tasks.

1 Introduction

Languages display a vast range of variation with respect to the form and meaning of definiteness. For example, while languages like English make use of definite and indefinite articles to distinguish between the discourse status of various entities (***the*** *car* vs. ***a*** *car* vs. *cars*), many other languages—including Czech, Indonesian, and Russian—do not have articles (although they do have demonstrative determiners).[1] Sometimes definiteness is marked with affixes or clitics, as in Arabic; sometimes it is expressed through other constructions, e.g. **?** shows that in Chinese (a language without articles), the existential construction is used to express indefinite subjects and the *ba-* construction for definite direct objects.

Aside from this variation in the form of (in)definite noun phrases (NPs) within and across languages, there is also variability in the semantic, pragmatic, and discourse–related functions expressed by (in)definites. We will refer to these as *communicative functions*[$^{\text{FT}}_{\text{A}}$ of (in)definiteness]. **?**, pp. 6–7, for instance, shows that the conditions under which an NP is marked as definite or indefinite (or not at all) are language-specific by contrasting English and French translations such as:

(1) He showed **extreme care**. (unmarked)
Il montra **un soin extrême**. (indef.)

(2) I love **artichokes** and asparagus. (unmarked)
J'aime **les artichauts** et les asperges. (def.)

(3) His brother became **a soldier**. (indef.)
Son frère est devenu **soldat**. (unmarked)

A cross-linguistic classification of communicative functions should be able to characterize the aspects of meaning that account for the different patterns of definiteness marking exhibited in (1–3): e.g., that (2) concerns a generic class of entities while (3) concerns a role filled by an individual. For more on communicative functions, see §2.

This paper describes a classifier that predicts communicative function labels for English NPs using lexical, morphological, and syntactic features. The contribution of our work is in both the output of the classifier and the model that it uses (features and weights). [$^{\text{A}}_{\text{B}}$ What about the Random forest classifier,

[1] Definite NPs, such as demonstratives, personal pronouns, and possessives are found in all languages.

- **NONANAPHORA** $[-A,-B]$
  - **UNIQUE** $[+U]$
    - **UNIQ_HEARER_OLD** $[-G,+O,+S]$
      - UNIQ_PHYSICAL_COPRESENCE $[+R]$ (13)
      - UNIQ_LARGER_SITUATION $[+R]$ (237)
      - UNIQ_PREDICATIVE_IDENTITY $[+P]$ (1)
    - UNIQ_HEARER_NEW $[-O]$ (36)
  - **NONUNIQUE** $[-U]$
    - **NONUNIQ_HEARER_OLD** $[+O]$
      - NONUNIQ_PHYSICAL_COPRESENCE $[-G,+R,+S]$ (39)
      - NONUNIQ_LARGER_SITUATION $[-G,+R,+S]$ (117)
      - NONUNIQ_PREDICATIVE_IDENTITY $[+P]$ (13)
    - NONUNIQ_HEARER_NEW_SPEC $[-G,-O,+R,+S]$ (231)
    - NONUNIQ_NONSPEC $[-G,-S]$ (181)
  - **GENERIC** $[+G,-R]$
    - GENERIC_KINDLEVEL (0)
    - GENERIC_INDIVIDUALLEVEL (131)
- **ANAPHORA** $[+A]$
  - **BASIC** $[+O,-B]$
    - SAME_HEAD (556)
    - DIFFERENT_HEAD (329)
  - **EXTENDED** $[+B]$
    - BRIDGING_NOMINAL $[-G,+R,+S]$ (43)
    - BRIDGING_EVENT $[+R,+S]$ (10)
    - BRIDGING_RESTRICTIVEMODIFIER $[-G,+S]$ (614)
    - BRIDGING_SUBTYPE_INSTANCE $[-G]$
    - BRIDGING_OTHERCONTEXT $[+O]$ (112)
- **MISCELLANEOUS** $[-R]$
  - PLEONASTIC $[-B,-P]$ (53)
  - QUANTIFIED (248)
  - PREDICATIVE_EQUATIVE_ROLE $[-B,+P]$ (58)
  - PART_OF_NONCOMPOSITIONAL_MWE (100)
  - MEASURE_NONREFERENTIAL (125)
  - OTHER_NONREFERENTIAL (148)

Figure 1: CFD (Communicative Functions of Definiteness) annotation scheme, with relative frequency in the corpus [$^{\text{NS}}_{\text{S}}$ TODO; maybe also show these for each attribute value]. Internal (non-leaf) labels are in bold; these are not annotated or predicted.[$^{\text{NS}}_{\text{S}}$ TODO: normalize capitalization] +/− values are shown for ternary attributes Anaphoric, Bridging, Generic, Hearer-Old, Predicative, Referential, Specific, and Unique; these are inherited from supercategories, but otherwise default to 0. Thus, for example, the full attribute specification for UNIQ_PHYSICAL_COPRESENCE is $[-A,-B,-G,+O,0P,+R,+S,+U]$.

we don't mention that?] The classifier predicts communicative function labels that capture aspects of discourse-newness, uniqueness, specificity, and so forth. Such functions are usually preserved in translation, even when the grammatical mechanisms for expressing them are different. Indeed, previous work has noted that machine translation systems face problems while translating from one language to another when the languages use different grammatical strategies (see §7). The communicative function labels also represent the discourse status of entities, making them relevant for entity tracking, knowledge base construction, and information extraction.

The model is a form-meaning mapping, consisting of the syntactic, lexical, and morphological features and weights that are predictive of communicative functions. This in itself is linguistically significant in that it shows the grammatical mechanisms beyond the articles *the* and *a* that are used for expressing definiteness in English.

To build our model, we leverage a cross-lingual definiteness annotation scheme (§2) and annotated English corpus (§3) from prior work (**?**). Our classifier, §4, is a supervised log-linear model akin to logistic regression, with features that combine lexical and morphosyntactic information with prespecified groupings of the communicative function labels[$^{\text{A}}_{\text{B}}$ again what about random forest?]; the evaluation measures (§5) include one that exploits these label groupings to award partial credit according to relatedness. [$^{\text{NS}}_{\text{S}}$ TODO: §6 obtain good performance? discover interesting features? §§5–7]

2 Annotation scheme

The literature on definiteness describes functions such as uniqueness, familiarity, identifiability, anaphoricity, specificity, and referentiality (**?????????????**, *inter alia*) as being related to definiteness.[2] For this work, we have adopted a scheme that is based on a combination of these functions, the annotation scheme for Communicative Functions of Definiteness (CFD), as described in **?**. It is summarized in fig. 1.

[2] The reductionist approaches to definiteness try to define it in terms of one or two of above mentioned communicative functions. For example, **?** proposes that the combination of uniqueness and a presupposition of familiarity underlie all definite descriptions. However, possessive definite descriptions (*John's daughter*) and the weak definites (*the son of Queen Juliana of the Netherlands*) are neither unique nor necessarily familiar to the listener before they are spoken. In contrast to the reductionist approaches are approaches to grammaticalization (**?**) in which grammar develops over time in such a way that each grammatical construction has some prototypical communicative functions, but may also have many non-prototypical communicative functions. The scheme we are adopting assumes that there may be multiple functions to definiteness.

Once upon a time there was a dear little girl who was loved by everyone who looked at her, but most of all by her grandmother, and there was nothing that she would not have given to the child.

Once she gave her a little riding hood of red velvet , which suited her so well that
SAME_HEAD DIFFERENT_HEAD OTHER_NONREFERENTIAL SAME_HEAD
NONUNIQ_HEARER_NEW_SPEC

she would never wear anything else; so she was always called 'Little Red Riding Hood.'
SAME_HEAD QUANTIFIED SAME_HEAD UNIQ_HEARER_NEW

Figure 2: An annotated sentence from "Little Red Riding Hood." The previous sentence is shown for context.

It was developed by annotating texts in two languages (English and Hindi) for various genres keeping in mind the communicative functions that have been associated with definiteness previously. The hierarchical organization of CFD serves to reduce the number of decisions that an annotator needs to make for speed and consistency.

It assigns a communicative function label to every NP except for first-person pronouns, second-person pronouns and relative pronouns. The three main communicative functions in CFD are **Anaphora** vs. **Nonanaphora** (whether the entity is old in the discourse or not), **Hearer-old** vs. **Hearer-new**, and **Unique** vs. **Nonunique** (annotated for **Nonanaphoric** only in the current scheme). However there are a few twists. Entities that have not been mentioned are considered **Anaphoric** (discourse-old) if they are evoked by a previously mentioned entity. For example, after mentioning a wedding, *the bride*, *the groom*, and *the cake* are considered to be **Anaphoric** (**??**). Entities that are **Non-Anaphoric** can be **Hearer-old** if they are physically present in the speech situation (**Physical-copresence**) or are not Physically-copresent but are identifiable by members of a community due to it being a part of the common knowledge retained by the community (**Larger-situation**, e.g., spoken on the first day of a conference: "I'm tired. The airplane was noisy.").[3]

In addition to the three main communicative functions, we have annotations for generic, pleonastic, quantified, predicative, and non-referential NPs. For details on the scheme, see **?**

Figure 2 is an excerpt from the "Little Red Riding Hood" annotated with the CFD scheme.

3 Data

We use the English definiteness corpus of **?**, which consists of texts from multiple genres annotated with the scheme described in §2. The 17 documents consist of prepared speeches (TED talks and a presidential address), published news articles, and fictional narratives. The TED data predominates (75% of the corpus)[4]; the presidential speech represents about 16%, fictional narratives 5%, and news articles 4%. All told, the corpus contains 13,860 words (868 sentences), with 3,422 NPs (the annotatable units). **?** report high inter-annotator agreement, estimating Cohen's $\kappa = 0.94$ within the TED genre and 0.91 for combined genres[$^{\text{A}}_{\text{B}}$ will enter the current IAA between Jordan & me soon].

4 Classification framework

To model the relationship between the grammar of definiteness and its communicative functions in a data-driven fashion, we work within the supervised framework of feature-rich discriminative classification, treating the functional categories from §2 as output labels y and various lexical, morphological, and syntactic characteristics of the language as features of the input x. Specifically, we learn a probabilistic log-linear model similar to multiclass logistic regression, but deviating in that logistic regression treats each output label (response) as atomic, whereas we decompose each into *attributes* based on their linguistic definitions, enabling commonalities between related labels to be recognized. Each weight in the model corresponds to a feature that mediates between *percepts* (characteristics of the input NP) and attributes (characteristics of the label). This is aimed at attaining better predictive accuracy as well as feature weights that better describe the form–function interactions we are interested in recovering.

Our setup is formalized below, where we discuss the mathematical model and linguistically motivated features.

[3] **?** also proposed a hierarchy with similar leaf nodes, but different internal structure.

[4] Note that the TED talks are from a large parallel corpus obtained from *http://www.ted.com/talks/*.

4.1 Model

At test time, we model the probability of semantic label y conditional on an [$^{\mathrm{NS}}_{\mathrm{S}}$ gold?] NP x as follows:

$$p_{\boldsymbol{\theta}}(y|x) = \log \frac{\exp \boldsymbol{\theta}^{\top}\mathbf{f}(x,y)}{\sum_{y' \in \mathcal{Y}} \exp \boldsymbol{\theta}^{\top}\mathbf{f}(x,y')} \tag{1}$$

[$^{\mathrm{CJ}}_{\mathrm{D}}$ Wouldn't it make more sense, given the story about percepts and label attributes, to use the form $\boldsymbol{\phi}(x)^T W \boldsymbol{\omega}(y)$? Then you could cast W as a linear map from the percept vector space to the attribute vector space. This is a bit different than the usual linear regression story (of course), but it's closer to what you're actually doing. I think the links to the feature vector function of x, y is a bit opaque and could be a footnote, but I'll leave it up to you.] where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a vector of parameters (feature weights), and $\mathbf{f} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is the feature function over input–label pairs. The feature function is defined as follows:

$$\mathbf{f}(x,y) = \boldsymbol{\phi}(x) \times \tilde{\boldsymbol{\omega}}(y) \tag{2}$$

where the percept function $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^c$ produces a vector of real-valued characteristics of the input, and the attribute function[$^{\mathrm{FT}}_{\mathrm{A}}$ define a and c to make it clear] $\tilde{\boldsymbol{\omega}} : \mathcal{Y} \to \{0,1\}^a$ encodes characteristics of each label. There is a feature for every percept–attribute pairing: so $d = c \cdot a$ and $f_{(i-1)a+j}(x,y) = \phi_i(x)\tilde{\omega}_j(y), 1 \le i \le c, 1 \le j \le a$. The contents of the percept and attribute functions are detailed in §4.2 and §4.3.

For prediction, having learned weights $\hat{\boldsymbol{\theta}}$ we choose the y that maximizes this probability [$^{\mathrm{CJ}}_{\mathrm{D}}$ maybe add: the Bayes-optimal decision rule for minimizing misclassification error]:

$$\hat{y} \leftarrow \arg\max_{y' \in \mathcal{Y}} p_{\hat{\boldsymbol{\theta}}}(y|x) \tag{3}$$

Training optimizes $\hat{\boldsymbol{\theta}}$ so as to maximize a convex L_1-regularized learning objective over the training data $\mathcal{D}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} -\lambda \|\boldsymbol{\theta}\|_1 + \sum_{\langle x,y \rangle \in \mathcal{D}} \log \frac{\exp \boldsymbol{\theta}^{\top}\mathbf{f}(x,y)}{\sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{\theta}^{\top}\mathbf{f}(x,y')\right)} \tag{4}$$

With $\tilde{\boldsymbol{\omega}}(y) =$ *the identity of the label*, this reduces to standard logistic regression.

4.2 Percepts

The characteristics of the input that are incorporated in the model, which we call *percepts* to distinguish them from model features linking inputs to outputs,[5] are intended to capture the aspects of English morphosyntax that may be relevant to the communicative functions of definiteness.

After preprocessing the text with a dependency parser and coreference resolver, we extract ~~the~~ several kinds of percepts for each NP.

4.2.1 Basic

Words of interest. These are the *head* within the NP, all of its *dependents*, and its *governor* (external to the NP). We are also interested in the *attached verb*, which is the first verb one encounters when traversing the dependency path upward from the head. For each of these words, we have separate percepts capturing: the token, the part-of-speech (POS) tag, the lemma, the dependency relation, and (for the head only) a binary indicator of plurality (determined from the POS tag). As there may be multiple dependents, we have additional features specific to the first and the last one. Moreover, to better capture tense, aspect and modality, we collect the attached verb's *auxiliaries*. We also make note of *neg* if it is attached to the verb.

Structural. The structural percepts are: the *path length* from the head up to the root, and to the attached verb. We also have percepts for the number of dependents, and the number of dependency relations that link non-neighbors. Integer values were binarized with thresholding.

[5] See above.[$^{\mathrm{FT}}_{\mathrm{A}}$ see section or subsection number?]

Positional. These percepts are the *token length* of the NP, the NP's *location* in the sentence (first or second half), and the *attached verb's position* relative to the head (left or right). 12 additional percept templates record the POS and lemma of the left and right neighbors of the head, governor, and attached verb.

4.2.2 Contextual NPs

When extracting features for a given NP (call it the "target"), we also consider NPs in the following relationship with the target NP: its *immediate parent*, which is the smallest NP whose span fully subsumes that of the target; the *immediate child*, which is the largest NP subsumed within the target; the *immediate precedent* and *immediate successor* within the sentence; and the *nearest preceding coreferent mention*.

For each of these related NPs, we include all of their basic percepts conjoined with the nature of the relation to the target.

4.3 Attributes

As noted above, though labels are organized into a tree hierarchy, there are actually several dimensions of commonality that suggest different groupings. These attributes are encoded as ternary characteristics; for each label (including internal labels), every one of the 8 attributes is assigned a value of +, −, or 0 (refer to fig. 1). In light of sparse data, we design features to exploit these similarities via the attribute vector function

$$\boldsymbol{\omega}(y) = [y, A(y), B(y), G(y), O(y), P(y), R(y), S(y), U(y)]^{\top}$$

where $A : \mathcal{Y} \to \{+,-,0\}$ returns the value for Anaphoric, $B(y)$ for Bridging, etc.$[^{\mathrm{FT}}_{\mathrm{A}}$ explain the three possible values: positive negative or 0] The identity of the label is also included in the vector so that different labels are always recognized as different by the attribute function. The categorical components of this vector are then binarized to form $\tilde{\boldsymbol{\omega}}(y)$; however, instead of a binary component that fires for the 0 value of each ternary attribute, there is a component that fires for *any* value of the attribute—a sort of bias term. The weights assigned to features incorporating + or − attribute values, then, are easily interpreted as deviations relative to the bias.

5 Evaluation

The following measures will be used to evaluate our predictor against the gold standard for the held-out evaluation (dev or test) set $\mathcal{E}$:

- **Exact match:** This accuracy measure gives credit only where the predicted and gold labels are identical.
- **By leaf label:** We also compute precision and recall of each leaf label to determine which categories are reliably predicted.
- **Soft match:** This accuracy measure gives partial credit where the predicted and gold labels are related. It is computed as the proportion of attributes whose (categorical) values match: $|\boldsymbol{\omega}(y) \cap \boldsymbol{\omega}(y')|/9$.$[^{\mathrm{CJ}}_{\mathrm{D}}$ Should these be differentially weighted based on their height in the hierarchy?]
- **Perplexity:** Perplexity is a value $\rho \in [1, \infty)$ that quantifies how "surprised", on average, our model is by the gold labels in the test set; the greater the probability mass assigned to the true labels, the lower the score. It is computed as $\rho = 2^{-(\sum_{\langle x,y\rangle \in \mathcal{E}} \log_2 p_{\hat{\boldsymbol{\theta}}}(y|x))/|\mathcal{E}|}$. Intuitively a perplexity of ρ means the amount of surprisal in a ρ-way uniform random draw.

6 Experiments

6.1 Experimental Setup

The annotated corpus of **?** (§3) contains 17 documents in 3 genres: 13 prepared speeches (mostly TED talks), 2 newspaper articles, and 2 fictional narratives. We arbitrarily choose some documents to hold out from each genre; the resulting test set consists of 2 TED talks ("Alisa_News", "RobertHammond_park"), 1 newspaper article ("crime1_iPad_E"), and 1 narrative ("Little Red Riding Hood"). The test set then contains 19,28 tokens (111 sentences), in which there are 511 annotated NPs; while the training set

Condition	$\lvert\boldsymbol{\theta}\rvert$	$\lVert\boldsymbol{\theta}\rVert_0$	λ	Exact Match Acc.	Soft Match Acc.	Perplexity
Majority baseline	—	—	—			
Log-linear classifier, no grouping by attributes						
Full log-linear classifier						

Table 1: Classifier versus baselines, as measured on the test set. The first three columns of numbers report the number of parameters (feature weights), the number of nonzero parameters, and the tuned regularization hyperparameter, respectively.

contains 2,911 NPs among 11,932 tokens (757 sentences). Gold NP boundaries are assumed throughout our experiments.

We use an in-house implementation of supervised learning with L_1-regularized AdaGrad (**?**). [$^{\mathrm{A}}_{\mathrm{B}}$ Now we mention L_2-regularization as well as or instead of L_1.] Hyperparameters are tuned on a dev set formed by holding out every tenth instance from the training set (test set experiments use the full training set).[$^{\mathrm{NS}}_{\mathrm{S}}$ early stopping? what exactly is tuned? optimize soft match acc?] Automatic dependency parses and coreference information were obtained with the parser and coreference resolution system in Stanford CoreNLP v. 3.3.0 (**??**) for use in features (§4.2).

6.2 Results

[$^{\mathrm{NS}}_{\mathrm{S}}$ English: ±cost function, ±non-identity attributes, ±predicting intermediate labels]

[$^{\mathrm{NS}}_{\mathrm{S}}$ maybe: which attribute groupings produce the best classifier, if we want to force a hierarchy]

[$^{\mathrm{NS}}_{\mathrm{S}}$ feature/attribute ablations]

[$^{\mathrm{A}}_{\mathrm{B}}$ a brief discussion of error analysis, feature weights-function mapping for possible grammaticalization??]

7 Related Work

[$^{\mathrm{A}}_{\mathrm{B}}$ am working on it currently, will update soon] [$^{\mathrm{NS}}_{\mathrm{S}}$ mention Bresnan's work on predicting syntactic alternations with logistic regression (here we want to predict the hidden information so that the classifier is useful for applications!).] Automated prediction of (in)definite articles has been found beneficial in a variety of applications, including postediting of MT output (**?**), text generation (**??**), and identification and correction of ESL errors (**????**).[$^{\mathrm{NS}}_{\mathrm{S}}$ what about coref? discourse models?] More recently, **?** trained a classifier to predict where English articles might plausibly be added or removed in a phrase, and used this classifier to improve the quality of statistical machine translation.

While definiteness morpheme prediction has been thoroughly studied in computational linguistics, studies on additional, more complex aspects of definiteness are limited. **?** exploit linguistically-motivated features in a supervised approach to distinguish between generic and specific NPs. To the best of our knowledge, no studies have been conducted on automatic prediction of semantic and pragmatic communicative functions of definiteness.

8 Conclusion

In future work we will build such models for languages that do not have articles such as Hindi, Russian, and Chinese.