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

Automatic Classification of Communicative Functions of Definiteness

Abstract

Whether an 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 supervised classifiers for English that use lexical, morphological, and syntactic features to predict communicative functions of definiteness. The benefits of this work are twofold: linguistically, the classifiers' features and parameters 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 classifiers may therefore be useful for coreference resolution, MT, and other mono- and cross-lingual 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 with other constructions, as in Chinese (a language without articles), where the existential construction is used to express indefinite subjects and the *ba*- construction to express definite direct objects (Chen, 2004).

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* of (in)definiteness. Croft (2003, 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 classifiers that predict communicative function labels for English NPs using lexical, morphological, and syntactic features. The contribution of our work is in both the output of the

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

- **NONANAPHORA** $[-A,-B]$ **999**
  - **UNIQUE** $[+U]$ **287**
    - **UNIQ_HEARER_OLD** $[-G,+O,+S]$ **251**
      - UNIQ_PHYSICAL_COPRESENCE $[+R]$ 13
      - UNIQ_LARGER_SITUATION $[+R]$ 237
      - UNIQ_PREDICATIVE_IDENTITY $[+P]$ 1
    - UNIQ_HEARER_NEW $[-O]$ 36
  - **NONUNIQUE** $[-U]$ **581**
    - **NONUNIQ_HEARER_OLD** $[+O]$ **169**
      - 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** $[+G,-R]$ **131**
    - GENERIC_KINDLEVEL 0
    - GENERIC_INDIVIDUALLEVEL 131
- **ANAPHORA** $[+A]$ **1574**
  - **BASIC** $[+O,-B]$ **795**
    - SAME_HEAD 556
    - DIFFERENT_HEAD 329
  - **EXTENDED** $[+B]$ **779**
    - BRIDGING_NOMINAL $[-G,+R,+S]$ 43
    - BRIDGING_EVENT $[+R,+S]$ 10
    - BRIDGING_RESTRICTIVEMODIFIER $[-G,+S]$ 614
    - BRIDGING_SUBTYPE_INSTANCE $[-G]$ 0
    - BRIDGING_OTHERCONTEXT $[+O]$ 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

	+	−	0		+	−	0		+	−	0		+	−	0
Anaphoric	1574	999	732	Generic	131	1476	1698	Predicative	72	53	3180	Specific	1305	181	1819
Bridging	779	1905	621	Hearer-Old	1327	267	1711	Referential	690	863	1752	Unique	287	581	2437

Figure 1: CFD (Communicative Functions of Definiteness) annotation scheme, with frequencies in the corpus. Internal (non-leaf) labels are in bold; these are not annotated or predicted. +/− 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 $[-A,-B,-G,+O,0P,+R,+S,+U]$. Counts for these attributes are shown in the table at bottom.

classifiers and the models themselves (features and weights). Each 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.

Our log-linear 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 models, we leverage a cross-lingual definiteness annotation scheme (§2) and annotated English corpus (§3) from prior work (Bhatia et al., 2014). Our classifiers, §4, are supervised models with features that combine lexical and morphosyntactic information with prespecified groupings of the communicative function labels; the evaluation measures (§5) include one that exploits these label groupings to award partial credit according to relatedness. §6 presents experiments comparing several models and discussing their strengths and weaknesses; computational work and applications related to definiteness are addressed in §7.

2 Annotation scheme

The literature on definiteness describes functions such as uniqueness, familiarity, identifiability, anaphoricity, specificity, and referentiality (Birner and Ward, 1994; Condoravdi, 1992; Evans, 1977, 1980; Gundel et al., 1988, 1993; Heim, 1990; Kadmon, 1987, 1990; Lyons, 1999; Prince, 1992; Roberts, 2003; Russell, 1905, *inter alia*) as being related to definiteness.[2] For this work, we have adopted a

[2]The reductionist approaches to definiteness try to define it in terms of one or two of above mentioned communicative functions. For example, Roberts (2003) 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 (Hopper and Traugott, 2003) 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

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 $\underbrace{\text{she}}_{\text{SAME\_HEAD}}$ gave $\underbrace{\text{her}}_{\text{DIFFERENT\_HEAD}}$ $\underbrace{\text{a little riding hood of } \underbrace{\text{red velvet}}_{\text{OTHER\_NONREFERENTIAL}}}_{\text{NONUNIQ\_HEARER\_NEW\_SPEC}}$, which suited $\underbrace{\text{her}}_{\text{SAME\_HEAD}}$ so well that $\underbrace{\text{she}}_{\text{SAME\_HEAD}}$ would never wear $\underbrace{\text{anything else}}_{\text{QUANTIFIED}}$; so $\underbrace{\text{she}}_{\text{SAME\_HEAD}}$ was always called $\underbrace{\text{'Little Red Riding Hood.'}}_{\text{UNIQ\_HEARER\_NEW}}$

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

scheme that is based on a combination of these functions, the annotation scheme for Communicative Functions of Definiteness (CFD), as described in Bhatia et al. (2014). It is summarized in fig. 1.

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** (Clark, 1977; Poesio and Vieira, 1998). 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 nonreferential NPs. Finally, because some natural groupings of labels would violate the hierarchy, each label type is associated with several attributes that can be +, −, or 0. For instance, with the Anaphoric attribute, a value of + applies to labels that can never mark NPs new to the discourse, − applies to labels that can *only* apply if the NP is new in the discourse, and 0 applies to labels such as PLEONASTIC (where anaphoricity is not applicable because there is no discourse referent). For details on the scheme, see Bhatia et al. (2014).

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 Bhatia et al. (2014), 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). Bhatia et al. (2014) report high inter-annotator agreement, estimating Cohen's $\kappa = 0.89$ within the TED genre as well as for all genres.

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 two kinds

definiteness.

[3]Komen (2013) 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/*.

of probabilistic models. The first is a 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. We also train a random forest model, which sacrifices interpretability of the learned parameters for predictive accuracy.

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

4.1 Models

We experiment with two classification methods: a log-linear model and a nonlinear tree-based ensemble model. Due to their consistency and interpretability, linear models are a valuable tool for quantifying and analyzing the effects of individual features. Non-linear models, while less interpretable, often outperform logistic regression (Perlich et al., 2003), and thus could be desirable when the predictions are needed for a downstream task.

4.1.1 Log-linear model

At test time, we model the probability of communicative function label y conditional on an NP x as follows:

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

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 $\tilde{\boldsymbol{\omega}} : \mathcal{Y} \to \{0,1\}^a$ 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 respectively.

For prediction, having learned weights $\hat{\boldsymbol{\theta}}$ we use the Bayes-optimal decision rule for minimizing misclassification error, selecting the y that maximizes this probability:

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

Training optimizes $\hat{\boldsymbol{\theta}}$ so as to maximize a convex L_2-regularized[5] learning objective over the training data $\mathcal{D}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} -\lambda ||\boldsymbol{\theta}||_2^2 + \sum_{\langle x,y \rangle \in \mathcal{D}} \log \frac{\exp \boldsymbol{\theta}^\top \mathbf{f}(x,y)}{\sum_{y' \in \mathcal{Y}} \exp \left(\boldsymbol{\theta}^\top \mathbf{f}(x,y') \right)} \tag{4}$$

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

4.1.2 Non-linear model

We employ a random forest classifier (Breiman, 2001), an ensemble of decision tree classifiers learned from many independent subsamples of the training data. Given an input, each tree classifier assigns a probability to each label; those probabilities are averaged to compute the probability distribution across the ensemble.

An important property of the random forests, in addition to being an effective tool in prediction, is their immunity to overfitting: as the number of trees increases, they produce a limiting value of the

[5] As is standard practice with these models, bias parameters (which capture the overall frequency of percepts/attributes) are excluded from regularization.

generalization error.[6] Thus, no hyperparameter tuning is required. Random forests are known to be robust to sparse data and to label imbalance, both of which are challenges with the definiteness dataset.

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,[7] 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 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.

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

$$\boldsymbol{\omega}(y) = [y, A(y), B(y), G(y), O(y), P(y), R(y), S(y), U(y)]^\top \tag{5}$$

where $A : \mathcal{Y} \to \{+, -, 0\}$ returns the value for Anaphoric, $B(y)$ for Bridging, etc. 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.

[6] See Theorem 1.2 in Breiman (2001) for details.

[7] See above.

Condition	$\lvert\boldsymbol{\theta}\rvert$	λ	Exact Match Acc.	Soft Match Acc.
Majority baseline	—	—	12.1	25.5
Log-linear classifier, attributes only	473,064	100	38.7	52.0
Log-linear classifier, labels only	413,931	100	40.8	52.1
Full log-linear classifier (labels + attributes)	926,417	100	43.7	55.6
Random forest classifier	20,363	—	**49.7**	**60.1**

Table 1: Classifiers and baseline, as measured on the test set. The first two columns give the number of parameters and the tuned regularization hyperparameter, respectively; the third and fourth columns give accuracies as percentages; the best in each column is bolded.

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: $|\boldsymbol{\omega}(y) \cap \boldsymbol{\omega}(y')|/9$.

6 Experiments

6.1 Experimental Setup

The annotated corpus of Bhatia et al. (2014) (§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 contains 2,911 NPs among 11,932 tokens (757 sentences). Gold NP boundaries are assumed throughout our experiments.

The log-linear model variants are trained with an in-house implementation of supervised learning with L_2-regularized AdaGrad (Duchi et al., 2011). Hyperparameters are tuned on a development set formed by holding out every tenth instance from the training set (test set experiments use the full training set): the power of 10 giving the highest soft match accuracy was chosen for λ.[8] The Python `scikit-learn` toolkit (Pedregosa et al., 2011) was used for the random forest classifier.[9] Automatic dependency parses and coreference information were obtained with the parser and coreference resolution system in Stanford CoreNLP v. 3.3.0 (Socher et al., 2013; Recasens et al., 2013) for use in features (§4.2).

6.2 Results

Measurements of overall classification performance appear in table 1. While far from perfect, our classifiers achieve promising accuracy levels given the small size of the training data and the number of labels in the annotation scheme. The random forest classifier is the most accurate, likely due to the robustness of that technique under conditions where the data are small and the frequencies of individual labels are imbalanced.

Among the log-linear models, the most successful is the richest model, which combines the fine-grained communicative function labels with higher-level attributes of those labels. This is encouraging because it suggests that the model has correctly exploited known linguistic generalizations to account for the grammaticalization of definiteness in English.

An advantage of log-linear models is that feature weights offer insights into the model's behavior. Figure 3 lists the 10 features that received the highest positive weights in the full model for the

[8]Preliminary experiments with cross-validation on the training data showed that the value of λ was stable across folds.

[9]Because it is a randomized algorithm, the results may vary slightly between runs; however, a cross-validation experiment on the training data found very little variance in accuracy.

+Specific	−Specific
`PRP$_first_pos_dependents`	`a_first_lemma_dependents`
`PRP$_pos_head_left`	`a_last_lemma_dependents`
`you_last_lemma_dependents`	`a_lemma_dependents_1`
`you_lemma_dependents_1`	`JJR_pos_head_left`
`PRP$_pos_dependents_1`	`JJR_last_pos_dependents`
`PRP$_pos_outerhead_right`	`a_lemma_dependents_2`
`NNP_last_pos_dependents`	`new_first_lemma_dependents`
`PRP$_last_pos_dependents`	`new_last_lemma_dependents`
`the_first_lemma_dependents`	`JJR_pos_dependents_2`
`from_lemma_outerhead`	`VB_pos_outerhead_left`

Figure 3: Percepts receiving highest positive weights in association with attribute values.

+ and − values of the Specific attribute. Reassuringly, the definite article, possessives (`PRP$`), proper nouns (`NNP`), and the second person pronoun are associated with specific NPs, while the indefinite article is associated with nonspecific NPs. The model also seems to have picked up on the less obvious but well-attested tendency of objects to be nonspecific (Aissen, 2003).[10]

In addition to confirming known grammaticalization patterns of definiteness, we can mine the highly-weighted features for new hypotheses: here the model thinks that objects of "from" are especially likely to be specific, and that NPs with comparative adjectives (`JJR`) are especially likely to be nonspecific. Whether these are general trends, or just an artifact of the sentences that happened to be in the training data, will require further investigation, ideally with additional datasets.

Finally, we can remove features to test their impact on predictive performance. Notably, in experiments ablating features indicating articles—the most obvious exponents of definiteness in English—we see a decrease in performance, but not a drastic one. This suggests that the expression of communicative functions of definiteness is in fact much richer than morphological definiteness.

Errors. Several labels are unattested or virtually unattested in the training data, so the models unsurprisingly fail to predict them correctly at test time. SAME_HEAD and DIFFERENT_HEAD, though both common, are confused quite frequently. Whether the previous coreferent mention has the same or different head is a simple distinction for humans; low model accuracy is likely due to errors propagated from coreference resolution. This problem is so frequent that merging these two categories and retraining the random forest model improves exact match accuracy by 8% absolute and soft match accuracy by 5% absolute. Another common confusion is between the highly frequent category UNIQUE_LARGER_SITUATION and the rarer category UNIQUE_HEARER_NEW; the latter is supposed to occur only for the first occurrence of a proper name referring to a entity that is not already part of the knowledge of the larger community. In other words, this distinction requires world knowledge about well-known entities, which could perhaps be mined from the Web or other sources.

7 Related Work

Because semantic/pragmatic analysis of referring expressions is important for many NLP tasks, a computational model of the communicative functions of definiteness has the potential to leverage diverse lexical and grammatical cues to facilitate deeper inferences about the meaning of linguistic input. We have used a coreference resolution system to extract features for modeling definiteness, but an alternative would be to predict definiteness functions as input to (or jointly with) the coreference task. Applications such as information extraction and dialogue processing could be expected to benefit not only from coreference information, but also from some of the semantic distinctions made in our framework, including specificity and genericity.

Better computational processing of definiteness in different languages stands to help machine translation systems. It has been noted that machine translation systems face problems when the source and the target language use different grammatical strategies to express the same information (Stymne, 2009;

[10]The percept `VB_pos_outerhead_left` fires when the NP is governed by a verb to its left.

Tsvetkov et al., 2013). Previous work on machine translation has attempted to deal with this in terms of either (a) preprocessing the source language to make it look more like the target language (Collins et al., 2005; Habash, 2007; Nießen and Ney, 2000; Stymne, 2009, *inter alia*); or (b) post-processing the machine translation output to match the target language, (e.g., Popović et al., 2006). Attempts have also been made to use syntax on the source and/or the target sides to capture the syntactic differences between languages (Liu et al., 2006; Yamada and Knight, 2002; Zhang et al., 2007). Automated prediction of (in)definite articles has been found beneficial in a variety of applications, including postediting of MT output (Knight and Chander, 1994), text generation (Elhadad, 1993; Minnen et al., 2000), and identification and correction of ESL errors (Han et al., 2006; De Felice and Pulman, 2008; Gamon et al., 2008; Rozovskaya and Roth, 2010). More recently, Tsvetkov et al. (2013) 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. Reiter and Frank (2010) 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 more broadly.

Our work is related to research in linguistics on the modeling of syntactic constructions such as dative shift and the expression of possession with "of" or "'s". Bresnan and Ford (2010) used logistic regression with semantic features to predict syntactic constructions. Although we are doing the opposite (using syntactic features to predict semantic categories), we share the assumption that reductionist approaches (as mentioned earlier) are not able to capture all the nuances of a linguistic phenomenon. Following Hopper and Traugott (2003) we observe that grammaticalization is accompanied by *function drift*, resulting in multiple communicative functions for each grammatical construction.

8 Conclusion

We have presented a data-driven approach to modeling the relationship between communicative functions associated with (in)definiteness and their lexical/grammatical realization in a particular language. Our feature-rich classifiers can give insight into this relationship as well as predict communicative functions for the benefit of NLP systems. This work has focused on English, but in future work we will build similar models for other languages—including languages without articles, under the hypothesis that such languages will rely on other, subtler devices to encode many of the functions of definiteness.

References

Judith Aissen. 2003. Differential object marking: iconicity vs. economy. *Natural Language & Linguistic Theory*, 21(3):435–483.

Archna Bhatia, Mandy Simons, Lori Levin, Yulia Tsvetkov, Chris Dyer, and Jordan Bender. 2014. A unified annotation scheme for the semantic/pragmatic components of definiteness. In *Proc. of LREC*. Reykjavík, Iceland.

Betty Birner and Gregory Ward. 1994. Uniqueness, familiarity and the definite article in English. In *Proc. of the Twentieth Annual Meeting of the Berkeley Linguistics Society*, pages 93–102.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.

Ping Chen. 2004. Identifiability and definiteness in Chinese. *Linguistics*, 42:1129–1184.

Herbert H. Clark. 1977. Bridging. In P.N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540. Association for Computational Linguistics, Ann Arbor, Michigan.

Cleo Condoravdi. 1992. Strong and weak novelty and familiarity. In *Proc. of SALT II*, pages 17–37.

William Croft. 2003. *Typology and Universals*. Cambridge University Press.

Rachele De Felice and Stephen G Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proc. of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 169–176. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Michael Elhadad. 1993. Generating argumentative judgment determiners. In *Proc. of AAAI*, pages 344–349.

Gareth Evans. 1977. Pronouns, quantifiers and relative clauses. *Canadian Journal of Philosophy*, 7(3):46.

Gareth Evans. 1980. Pronouns. *Linguistic Inquiry*, 11.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proc. of IJCNLP*, volume 8, pages 449–456.

J.K. Gundel, N. Hedberg, and R. Zacharski. 1988. The generation and interpretation of demonstrative expressions. In *Proc. of XIIth International Conference on Computational Linguistics*, pages 216–221.

J.K. Gundel, N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. In *MT Summit XI*, pages 215–222. Copenhagen.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers.

I. Heim. 1990. E-type pronouns and donkey anaphora. *Linguistics and Philosophy*, 13:137–177.

Paul J. Hopper and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge University Press.

N. Kadmon. 1987. *On unique and non-unique reference and asymmetric quantification*. Ph.D. thesis, University of Massachusetts.

N. Kadmon. 1990. Uniqueness. *Linguistics and Philosophy*, 13:273–324.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *Proc. of the National Conference on Artificial Intelligence*, pages 779–779. Seattle, WA.

Erwin Ronald Komen. 2013. *Finding focus: a study of the historical development of focus in English*. LOT, Utrecht.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 609–616. Association for Computational Linguistics, Sydney, Australia.

C. Lyons. 1999. *Definiteness*. Cambridge University Press.

Guido Minnen, Francis Bond, and Ann Copestake. 2000. Memory-based learning for article generation. In *Proc. of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language* *learning-Volume 7*, pages 43–48. Association for Computational Linguistics.

Sonja Nießen and Hermann Ney. 2000. Improving smt quality with morpho-syntactic analysis. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 1081–1085. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Claudia Perlich, Foster Provost, and Jeffrey S. Simonoff. 2003. Tree induction vs. logistic regression: a learning-curve analysis. *Journal of Machine Learning Research*, 4:211–255.

M. Poesio and R. Vieira. 1998. A corpus based investigation of definite description use. *Computational Linguistics*, 24:183–216.

Maja Popović, Daniel Stein, and Hermann Ney. 2006. Statistical machine translation of german compound words. In *Advances in Natural Language Processing*, pages 616–624. Springer.

E.F. Prince. 1992. The zpg letter: Subjects, definiteness and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: identifying singleton mentions. In *Proc. of NAACL-HLT*, pages 627–633. Atlanta, Georgia, USA.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proc. of ACL*, pages 40–49. Uppsala, Sweden.

C. Roberts. 2003. Uniqueness in definite noun phrases. *Linguistics and Philosophy*, 26:287–350.

Alla Rozovskaya and Dan Roth. 2010. Training paradigms for correcting errors in grammar and usage. In *Proc. of NAACL-HLT*, pages 154–162. Association for Computational Linguistics.

B. Russell. 1905. On denoting. *Mind, New Series*, 14:479–493.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proc. of ACL*, pages 455–465. Sofia, Bulgaria.

S. Stymne. 2009. Definite noun phrases in statistical machine translation into Danish. In *Proc. of Workshop on Extracting and Using Constructions in NLP*, pages 4–9.

Yulia Tsvetkov, Chris Dyer, Lori Levi, and Archna Bhatia. 2013. Generating English determiners in phrase-based translation with synthetic translation options. In *Proc. of WMT*. Association for Computational Linguistics.

Kenji Yamada and Kevin Knight. 2002. A decoder for syntax-based statistical mt. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 303–310. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA.

Yuqi Zhang, Richard Zens, and Hermann Ney. 2007. Improved chunk-level reordering for statistical machine translation. In *IWSLT 2007: International Workshop on Spoken Language Translation*, pages 21–28.