Title, author list not final. Page limit: 5 + 2 bib

[A B Learning Definiteness across Languages using Classifiers] [NS S maybe: A Classification and Classifier for Cross-linguistic Functions of Definiteness]

Abstract

Definiteness is a nonhomogeneous category across languages expressing various communicative functions (types of semantic, pragmatic and discourse- related information). Languages differ in the grainsize and inventory of communicative functions that are expressed as well as the syntactic means for expressing them. There is a major difference between languages that express definiteness primarily through articles like English "the" and "a" and languages that express definiteness with word order and other syntactic constructions. This paper presents an [Y T a thorough linguistic analysis of] annotation scheme for the communicative functions of definiteness. [Y T Agree with Nathan – the scheme was presented in the LREC paper. This paper presents the linguistic analysis and the classifier] We applied the annotation scheme to a corpus of written and spoken English and built a classifier that uses lexical, morphological, and syntactic features to predict communicative functions. This work has twofold benefits: linguistically, it discovers the grammaticalization of definiteness, some of which is obvious (e.g., articles) and some of which is not obvious (include an example from our findings) In addition, this work has the potential to improve language technologies such as machine translation, information extraction, entity tracking, and other semantic processing tasks. [Y T We work only on English, and I don't think we discover grammaticalization. I'd say that benefits are: (1) linguistically, it provides analysis and interpretation of the scheme, (2) computationally, it presents a framework to predict *semantic and pragmatic* properties of the communicative functions of definiteness which, unlike lexical and morphosyntactic features, are preserved in translation. This framework can be used to discover definiteness grammaticalization across languages.].

1 Introduction

Languages display a vast range of variation with respect to 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*), many other languages, such as Czech, Hindi, Indonesian, Russian do not have definite articles or any articles at all. Even the languages that do have articles vary in terms of how these articles are used. For example, English uses only one definite article 'the' for entities irrespective of whether they have been mentioned previously in the discourse or are familiar to the hearer by some other means, but there are languages such as Hausa which have a specialized article for anaphoric definite entities, see example in Figure **??** (borrowed from **?**).

```
(1) Na  kawo  keke    din-ka
    AUX bring bicycle the-2SGM
    'I brought your bicycle (previously mentioned).'
```

There are other means to express (in-)definiteness, such as the use of affixes. Arabic, for example, uses a definite prefix *al-* and indefinite suffix *-n*. **?** shows that Chinese expresses (in-)definiteness through the use of various constructions, such as the existential construction for indefinite subjects and the *ba-* construction for definite direct objects. Demonstratives, personal pronouns and possessives (which are found in all languages) are other kinds of definite NPs. Besides this variation in the form of (in-)definite NPs within and across languages, there is also variability in what semantic, pragmatic, discourse-

related functions (in-)definites express. We will refer to these as communicative functions.

The literature on definiteness suggests that it may express communicative functions corresponding to notions such as uniqueness, familiarity, identifiability, anaphoricity, specificity, and referentiality(**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?** among many others). Reductionist approaches to definiteness try to define a single communicative function of definteness. For example, **?**, **?** have used the notion of semantic uniqueness to account for definite NPs, but pronouns do not display uniqueness effects. **?** proposes that the combination of uniqueness and a presupposition of familiarity can explain definiteness, but possessive definite descriptions (*John's daughter*) and the weak definites (*the son of Queen Juliana of the Netherlands*) cannot be explained using this combination.

In contrast to the reductionists, we ... (I'll finish editing this paragraph tomorrow. It will say that we have a range of prototypical and radial communicative functions and we want to correlate them with lexical, syntactic, and morphological constructions.)

We take such linguistic observations to suggest that definiteness is not as homogeneous a category as many accounts have assumed. Instead, it should be seen as grammaticalization of many such semantic, pragmatic, discourse- related notions. Therefore, it seems relevant to identify what information different forms of (in-)definites express within a language as a goal in itself. Additionally, it is advantageous to identify the mapping between the form and meaning of (in-)definites for various NLP applications. For example, this mapping can be used to encode correct information during the automatic construction of knowledge bases or other information extraction or semantic processing tasks. Also, this mapping can be used for machine translation applications. Here the assumption is that even though the forms may change across languages, the meaning is maintained while translating from one language to another. It has been noted previously that machine translation systems face problems while translating from a language to another when the languages use different grammatical strategies. **?** and **?** mention how translating from an article-language to an article-less language is problematic. If the mapping between different forms and the information they encode is known in both the source language and the target language, this information can be leveraged in improving machine translation across these languages.

In §2, we review the annotation scheme we have used to determine the mapping between the form and meaning it encodes with respect to (in-)definiteness. In §3, we provide information about the English data we have annotated using this scheme. §4 is an overview of the classification framework we have used. In §5, we describe the measures we have used to evaluate the predictions against the gold standard for the held-out data. In §6, we describe our experiments and discuss the results. In §7, we summarize our findings and briefly discuss the future tasks we are undertaking or envision. [$^{Y}_{T}$ I think that detailed description of definiteness in linguistics in the beginning of this section deserves a section by itself; I'd split the Introduction section into Introduction (shorter, only what we do, motivation, contribution) and then Definiteness.]

2 Annotation scheme

Based on the literature on definiteness, we compiled a list of semantic, pragmatic, and discourse-related features which are relevant for the form of (in-)definite NPs in English (and many other languages). [$^{Y}_{T}$ ref to the LREC paper?] These features were put together in a hierarchical structure to ease the annotation effort for the annotators. At each step, the annotators make a decision and thus reduce the number of labels they have to consider before annotating an NP. This approach also increases the consistency in annotations due to the reduced effort on considering the options among which to select the correct label for the NP. The annotation scheme was revised through many iterations after attempts at annotating natural language data from different genres using the scheme. The current scheme is presented in Figure 1.

The first distinction annotators make is whether the NP at hand is anaphoric (discourse old) or non anaphoric (discourse new). We do not annotate for uniqueness, hearer old/new distinction, specificity or genericity for the anaphoric NPs as those distinctions were deemed inconsequential for English anaphoric NPs[1]. [$^{Y}_{T}$ For someone who is not familiar with the scheme, this would be very hard to understand without looking at the definiteness

[1]However, in future, we may include these features for anaphoric NPs as well if contrary evidence were found.

scheme. I think the scheme should be on this page, and description should always point to the place in the scheme. May be we can write all scheme-related terms using the same font as they appear in the scheme?] The anaphoric NPs' category is further subdivided into basic anaphora and extended anaphora. The basic anaphora is used for entities which have been mentioned previously in the discourse. The extended anaphora is inspired by the bridging references **?** had mentioned. These are used for entities which are not discourse old or even hearer old, but at the same time, they are not entirely new either. See example in Figure **??** (borrowed from **?**).

(2) I went to *a wedding* last weekend. *The bride* was a friend of mine. She baked *the cake* herself.

Within the non anaphoric category, the annotators decide whether the NP denotes a unique NP, a non unique NP or a generic NP. For the non-generic cases, further decisions are made based on whether the entity is hearer old or new, and if it is old, what kind of situation makes it old. For generics also, there are two categories depending on whether they appear with kind-level or individual-level predicates resulting in different properties of these NPs. Besides these, there are a few non-referential categories as well.

Besides the annotation categories, a few other decisions were also taken regarding annotations. For example, corresponding to the basic anaphora cases, the anaphoric links can be established with antecedent NPs in previous sentences as well as with NPs within the same sentence. Also to determine whether an NP gets a Same_head or a Different_head label, we refer to the closest antecedent even if it appears within the same sentence. Regarding annotations of nested NPs, we assume that these NPs are interpreted inside out. This helps determine the cases of Bridging_RestrictiveModifier category. If an NP consists of a modifier that restricts its reference enough to make the referent identifiable to the addressees, the embedding NP is annotated with the Bridging_RestrictiveModifier label.

Figure **??** is an excerpt from Little Red Riding Hood demonstrating the above-mentioned decisions regarding annotations and illustrating some of the categories from our annotation scheme.

. . . . . .

[$^{NS}_{S}$ we should probably have a name for the scheme—something like "Functions of Definiteness Across Languages (FDAL)"? and a name for the categories (I have been calling them semantic functions).][$^{A}_{B}$ I actually like the name "Functions of Definiteness Across Languages" but am not sure how to incorporate it here.] [$^{NS}_{S}$ suggestion: lead with a discourse excerpt, ideally illustrating a few of the categories, a nested NP, and an anaphoric link across sentence boundaries. I can help format it.]

[$^{NS}_{S}$ cite papers like this: **?** (but probably anonymize this for review) or (**?**)]

3 Data

We annotated data from a few genres, namely TED talks, news articles, speech (Presidential inauguration speech) and stories in English. However, currently the majority of our annotated data is from the TED talks corpus. For future work, we will be annotating other genres more as well. We annotated a total of 1783 sentences [$^{A}_{B}$ that is all our data, some is not yet annotated, am working on it today, hopefully we will be able to use most of it, otherwise we will change the numbers here] (a total of 28574 words), which consist of 10476 NPs (i.e. the annotatable units).

A total of 6 annotators contributed to the annotations. Two of the annotators are linguists while others are computer scientists working on languages and an undergraduate student of Linguistics. Two annotators annotated most of the data, the inter annotator agreement between them was very high with Cohen's Kappa score of 0.94 within the TED corpus and 0.91 for combined genres. However the other 4 annotators also annotated some of the data, but one of the two annotators with most experience took these data as inputs and revised annotations according to their annotations. This two-step process of annotation was taken to increase the amount of data within a short period of time and also to achieve consistency within annotations. All annotators were trained using the annotation guidelines and example annotated texts. Then the annotators were asked to annotate some data, these annotations were discussed to reach at consensus. The annotations used for discussions to reach at consensus were not included while calculating the inter annotator agreement.

[$^{A}_{B}$ description of data, IAA, other statistics-n(annotations), n(annotations/ category) classi-

- **NONANAPHORA** $[-A,-B]$ **(00)**
  - **UNIQUE** $[+U]$ **(00)**
    - ∗ **UNIQ_HEARER_OLD** $[-G,+O,+S]$ **(00)**
      - · UNIQ_PHYSICAL_COPRESENCE $[+R]$ (00)
      - · UNIQ_LARGER_SITUATION $[+R]$ (00)
      - · UNIQ_PREDICATIVE_IDENTITY $[+P]$ (00)
    - ∗ UNIQ_HEARER_NEW $[-O]$
  - **NONUNIQUE** $[-U]$
    - ∗ **NONUNIQ_HEARER_OLD** $[+O]$
      - · NONUNIQ_PHYSICAL_COPRESENCE $[-G,+R,+S]$
      - · NONUNIQ_LARGER_SITUATION $[-G,+R,+S]$
      - · NONUNIQ_PREDICATIVE_IDENTITY $[+P]$
    - ∗ NONUNIQ_HEARER_NEW_SPEC $[-G,-O,+R,+S]$
    - ∗ NONUNIQ_NONSPEC $[-G,-S]$
  - **GENERIC** $[+G,-R]$
    - ∗ GENERIC_KINDLEVEL
    - ∗ GENERIC_INDIVIDUALLEVEL
- **ANAPHORA** $[+A]$
  - **BASIC** $[+O,-B]$
    - ∗ SAME_HEAD
    - ∗ DIFFERENT_HEAD
  - **EXTENDED** $[+B]$
    - ∗ BRIDGING_NOMINAL $[-G,+R,+S]$
    - ∗ BRIDGING_EVENT $[+R,+S]$
    - ∗ BRIDGING_RESTRICTIVEMODIFIER $[-G,+S]$
    - ∗ BRIDGING_SUBTYPE_INSTANCE $[-G]$
    - ∗ BRIDGING_OTHERCONTEXT $[+O]$
- **MISCELLANEOUS** $[-R]$
  - PLEONASTIC $[-B,-P]$
  - QUANTIFIED
  - PREDICATIVE_EQUATIVE_ROLE $[-B,+P]$
  - PART_OF_NONCOMPOSITIONAL_MWE
  - MEASURE_NONREFERENTIAL
  - OTHER_NONREFERENTIAL

Figure 1: Taxonomy of definiteness functions, with number of occurrences in the training data [NS S TODO]. Internal (non-leaf) labels are in bold.[NS S added two: Generic and Miscellaneous, and also reversed Nonreferential to Referential. is that OK?][A B yes, that is fine][NS 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.[NS S nonuniq_nonspec is −generic, right?] [A B right] Thus, for example, the full attribute specification for UNIQ_PHYSICAL_COPRESENCE is $[-A,-B,-G,+O,0P,+R,+S,+U]$.

fiers for English- baseline, accuracy etc, confusion matrix- classifier predictions and gold standard may be a brief discussion of feature ablation study- which features are most helpful at least in these languages corresponding to individual categories (a clue to grammaticalization) Principle component analysis, clustering of features]

4 Classification framework

To model the relationship between the grammar of definiteness and its semantic 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 the following ways:

- Logistic regression treats each output label (response) as atomic; 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 noun phrase) and attributes (characteristics of the label).
- Logistic regression assumes a prediction is either correct or incorrect. We incorporate a *cost function* that gives partial credit during learning when a related label is predicted, so the learned model will better match our evaluation measure.
- Logistic regression assumes the space of possible predictions matches the space of labels observed in the training data; we allow more abstract labels to be predicted, which can receive partial credit. The scoring scheme encourages the predictor to "back off" to a coarser label if it is not sufficiently confident about a fine-grained label.

These decisions are 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.

4.1 Model

At test time, we model the probability of semantic label y conditional on a [NS S gold?] noun phrase x as follows:

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

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 func-

tion 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 §§4.3 and 4.2.

For prediction, having learned weights $\hat{\boldsymbol{\theta}}$ we choose 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 the L_1-regularized *softmax-margin* learning objective (**?**) over the training data $\mathcal{D}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathcal{D})$$

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

The *cost function* allows us to penalize some errors more than others during training, taking into account the linguistic functions of the labels. It is zero for the gold label and nonnegative for the others.[NS intuition]

This framework gives us several ways to design a classifier appropriate to the task: the attributes, the space of labels $\mathcal{Y}$ to consider, and the cost function, and of course, the features themselves. With $cost(y,y') = 0$, $\mathcal{Y} = \{$*gold labels in training*$\}$, and $\tilde{\boldsymbol{\omega}}(y) =$ *the identity of the label*, this reduces to standard logistic regression.

4.2 Features

Features of words of interest The text is parsed with a dependency parser. In an NP parse there are certain words that we are mostly interested in. They include its *head*, *dependents*, and *outer heads* (which are words that the heads depend on). We are also interested in the *attached verb*, which is the first verb one encounters when going upward the dependency path from the head. We extract their part-of-speech, lemmas, and grammatical information (of which we only collect plurality right now). Since there may be multiple dependents, we have special features for the first and the last one. Moreover, to better capture tense, aspect and modality, we collect the attached verb's *aux* words. We also make note of *neg* if it is attached to the verb. Of these words we collect their lemmas, POS's and dependency relation. Since there may be many dependents, we also specifically mark the first and the last word.

Parse structure related features These features include length of the upward path from the head to the root, and to the attached verb. We also have features for the number of dependents, and the number of dependency relations that link non-neighbors.

Position related features Including the length of the NP, NP position in the sentence (upper half or lower half), relative position of the verb to the head (left or right). We also collect information of the words of interest' left and right neighbors.

Other NPs These include features in annotated NPs such as *immediate parent*, which is the smallest NP that fully covers this NP, and similarly *immediate child*. We also have *immediate precedent* and *immediate successor*. Using a coreference resolution program, we identify mentions, and extract their features as well.

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 order to capture these similarities in the model's features and cost function, we define 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 $A : \mathcal{L} \cup \mathcal{I} \to \{+,-,0\}$ 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.

4.4 Cost and label space

The definiteness function hierarchy presented in fig. 1 consists of 24 *leaf labels*, which will be denoted $\mathcal{L}$, and 10 more abstract *intermediate labels*, denoted $\mathcal{I}$. All of the gold labels in the training data are from $\mathcal{L}$, but we give our model the option to predict more abstract labels to receive partial credit. We will therefore use $\mathcal{Y} = \mathcal{L} \cup \mathcal{I}$.

The relatedness of leaf label pairs is determined by the dot product of the two original attribute vectors: let $\Delta(\ell, \ell') = |\boldsymbol{\omega}(\ell) \cap \boldsymbol{\omega}(\ell')|^{-1}$.[2] For a gold leaf label ℓ and an internal label ι, $\Delta(\ell, \iota)$ = the distance in the hierarchy between ℓ and ι.[$^{\text{NS}}_{\text{S}}$ TODO: ensure if ι is an ancestor of ℓ, this is less than choosing another leaf dominated by ι]. (There is no need to define $\Delta(\iota, \cdot)$, as the training set does not contain intermediate labels.)

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 gives credit only where the predicted and gold labels are identical. When the model is allowed to predict internal labels, we will report overall precision and recall of leaf labels. Otherwise, we report accuracy.
- **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 Δ function in §4.4 [$^{\text{NS}}_{\text{S}}$ normalized to be between 0 and 1?].
- **Perplexity:** This determines how "surprised" our model is by the gold labels in the test set; the greater the probability mass assigned to the true labels, the higher the score. It is computed as $2^{\left(\Sigma_{\langle x,y\rangle \in \mathcal{E}} \log_2 p_{\hat{\boldsymbol{\theta}}}(y|x)\right)/|\mathcal{E}|}$.

[2] By the intersection of two attribute vectors, we mean the subset of components that have a matching (categorical) value.

6 Experiments

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

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

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

[$^{\text{NS}}_{\text{S}}$ Hindi?]

7 Conclusion

References