Determining Definiteness from Context

Christopher Brown
University of Texas at Austin
chrisbrown@utexas.edu

Abstract

Noun phrases occurring in natural language can be described as either definite or indefinite. Definiteness is a semantic concept that relates to givenness, familiarity, topicality, and focus. It can be denoted by a variety of determiners, the absence of a determiner, or by context. Focusing on the definite markers, *a* and *the*, this paper investigates what types of models are best at predicting the most suitable determiner when it is not available. This inference is useful in natural language generation and some cases of machine translation.

1 Problem and Motivation

Definiteness is an attribute of a noun phrase that describes the nature of its intended referent.

(1) **A dog** was barking all night.

(2) **Dogs** were barking all night.

(3) **This dog** was barking all night.

(4) **The dog** was barking all night.

In each case, the bold noun phrase picks out a different set of dogs or implies something different about which dog the speaker means to refer to. (1) is the least definite; (2) is similar, but allows the dogs to take shifts; (3) suggests the speaker has some way of uniquely identifying the dog, but the hearer would not know which dog without more information; (4) requires both hearer and speaker to know which dog is referred to.

Definiteness is generally defined as the form of noun phrase that is acceptable when the uniqueness of referent is known by both speaker and hearer. Specificity differs in that the hearer may not be aware of this uniqueness. Thus (3) is specific, but not definite.

In English, determiners are the primary overt markers of definiteness / indefiniteness. Definiteness is not completely specified by determiners, though; some uses of *a* can be definite ("A man just proposed to me in the orangery."[1]); *the* can be indefinite ("After entering the lobby, take the elevator to your left"—even though there may be multiple elevators), while *this* and bare plurals are much more ambiguous. There is a wide variety of other determiners and quantifiers, as well as most pronouns, which mark definiteness in some way.

1.1 Ubiquity

Determiners are very common in the English language (Table 1); like many other functional words, determiners are often redundant—misuse is very quickly detected, and the intended use can usually be determined from context. Many studies of definiteness approach the problem via language acquisition; misuse of determiners by learners of languages with definite-marking articles are observed to demonstrate significant patterns, which I discuss later in the section devoted to related work.

1.2 Simplifying the problem

I could not find any large corpora annotated for definiteness; Calhoun et al. (2010) describes a corpus where 'Information Status' is encoded, but not specifically definiteness. Vieira and Poesio (2000) developed a corpus of about $1,400$ noun phrases annotated with definiteness, but this is relatively small, and the available annotations are relatively deep. I

[1] Fodor and Sag, 'Referential and Quantificational Indefinites', *Linguistics and Philosophy* 5. 1982.

Token	Count	Percentage
the	69968	6.83%
of		
and		
to		
a	23490	2.29%
in		
that	10786	1.05%
...		
this	5145	0.50%
...		
an	3746	0.36%

Table 1: Brown corpus counts and proportions of selected definiteness markers (in bold).

am unaware of any attempts to model definiteness computationally; definiteness is discussed at length in linguistic literature, but its limited usefulness in popular NLP tasks has left it without much empirical / computational attention.

Definiteness effects overlap with other semantic features, particularly presupposition and the scope of context. As with many linguistic aspects of semantics / pragmatics, a complete solution would require a much fuller understanding of language than any current NLP system provides.

In many cases, definiteness is clear from the determiner, as with *a* and *the*. Noun phrases with *a* (and *an*) are predominantly indefinite, and those with *the* are always definite. The analysis described in this paper assumes that these particular determiners cover enough cases to serve as a significant portion of the problem, and that *a*/*an* always denote indefiniteness, and *the* always denotes definiteness. The naive baseline accuracy starts pretty high: 69.1%—from a maximum likelihood estimate which simply uses the more common of the two markers (*the*). This paper demonstrates an approach that achieves up to 79.9% accuracy using SVM (support vector machines) and a small selection of well-motivated, intuitive features. I show that it's possible to recover information about definiteness from nearby nouns and similarly shallow features. I do this by replacing all determiners *a*, *an*, and *the* with a special placeholder token, using the original form as the gold label (only distinguishing definite from indefinite). Additionally, I compare results of CRF (Conditional Random Field) sequence labeling against a SVM (Support Vector Machine) tagging each token alone. This is primarily a feature selection problem, and I focus on results comparing different subsets of features.

2 Applications

Detecting or inferring definiteness fills a very specific niche of natural language processing. Determiners serve as useful features in POS (Part of Speech) tagging, since they usually mark the beginning of a noun phrase, and in NER (Named Entity Recognition), for the same reason. Such methods regard determiners as a side-effect, though, and don't necessarily distinguish one determiner from another or seek to understand the determiners as independent linguistic features.

This linguistic analysis of understanding determiners at the level of definiteness primarily contributes to two common NLP tasks: language generation and machine translation.

2.1 Language generation

When generating language from a logical structure, co-referential noun phrases representing a single entity must surface fluently. It is unnatural to repeat a proper noun or an extended noun phrase when a pronoun would suffice. But context and focus shift as the discourse develops, eventually making any pronoun ambiguous, forcing the speaker to use a more explicit reference. Tracking this throughout a discourse requires understanding the shared information between hearer and speaker, which naturally shifts over time throughout the discourse.

This is often modeled with centering theory—of which there are many variations, but they all involve some kind of stage that only supports or holds a limited number of entities. When generating natural language from the logical form of a series of events, we must track what alternatives are available at each point, in order to know what level of definiteness is most appropriate, which in turn will determine what sort of noun phrase we use—definite, indefinite, pronoun, etc. Syntactic binding also plays a role here, in determining whether we use a pronoun or a (reflexive) anaphor.

Definiteness is just one part of anaphora generation, and this paper tackles only a partial, but integral, part of the problem. The aspect that is relevant here requires shallower understanding. If we use a full noun phrase more than once, we need to know when to use an indefinite description and when to use the definite form.

2.2 Machine translation

Some languages, like Russian and Korean, do not have determiners to denote definiteness; others, like Japanese and Hindi, have a few articles that convey some semantic level of indefiniteness, usually, or are used for other special purposes (such as to denote humans). When translating between these languages and languages with overt definiteness markers, we must insert determiners that often have no parallel marker in the source text.

For example, Google Translate renders each of the four alternations of "A/The man bit a/the dog" as Человек укусил собаку. Translating back into English produces "A person bitten by a dog." I don't know Russian; it may be genuinely ambiguous, and only determinable via context; but I know that translating back into English accurately requires choosing the correct determiners, for which we need a model of definiteness (as well as anaphora / centering).

There are a few papers on the second of these applications, translation, but they are specific to only a few language pairs (Ishikawa, 1995, Siegel, 1996).

2.3 General contribution

This can explain why so little computational work exists on the subject, but the linguistic and philosophical literature on definiteness is much more developed. However

These applications are useful to two specific niches of NLP, , and apply only in some cases of those areas.

3 Experiment

The basic research question will be to determine what type of language models produce the best prediction of definiteness. Because English has determiners, any parsed document is "labeled" data. I will simply replace *a* and *the*'s with the placeholder «DET», retaining the original token as the label for that token. While this is an artificial evaluation metric, it could be useful in both of the application scenarios. In anaphora generation, each instance has to be filled by either a full noun phrase, either definite or indefinite, or a pronoun. In the noun phrase case, a placeholder would be inserted by a first pass of the language generation model, and then resolved by the anaphora resolution pass. In machine translation, noun phrases translated from a determiner-less language to one with determiners could be padded with a placeholder, which would then be resolved by some post-processing step performed on just the target language.

My project will consist of evaluating different models for predicting the deleted marker; it is not a typical sequence labeling problem, since relatively few of the tokens need to be resolved. But neither is each instance independent of the previous; in fact, a prior mention is presumably a strong indicator that an entity is now in the active context, and subsequent instances should use the definite determiner. The sequence is crucial in a sense of building and maintaining a 'center,' which is a common approach when resolving anaphora (Grosz et al., 1995, Beaver, 2000). But with an intelligent set of features, we might achieve sufficient accuracy with a simple logistic regression; given a set of features, evaluate whether the placeholder should be definite or indefinite?

My investigation will also compare feature sets, to determine which are the most efficient at predicting the deleted definiteness markers. Is the parse structure important, or just the surrounding tokens? Position in sentence / document? One presumably useful feature would be to use a cumulative index for each token type in the document—1 for the first instance, 2 for the second, and so on—which might handle the prior mention issue without requiring a sequence model.

4

5 Semantics of definiteness

The theoretical semantics of definite noun phrases has an extensive history; we can start with Gottlob Frege and Bertrand Russell and names. Frege said that the meaning of a name (or definite noun phrase) was the object denoted by that description. Russell complicated matters with his “The King of France is bald” example, in which “the King of France” cannot refer to anything (at least not since 1848). Russell would call this sentence false, despite the description’s failure to refer to anything; the failure of the noun phrase propagates out to the failure of the sentence.

Keith Donnellan and P. F. Strawson disagree; they say that referencing is something speakers, not words, do, and that “the King of France” is simply unsuccessful. Compare “a King of France is bald,” which can refer to any (dead) past King of France, and thus have a truth value. Definite noun phrase usually denote a unique entity, and if this fails, the sentence is incoherent. But compare “After you enter the lobby, take the elevator to the 13th floor,” which is acceptable even if there is more than one elevator.

This is just the beginning, but it’s clear that the difference between definite and indefinite noun phrases involves a number of factors, such as context, real world knowledge, and other pragmatic phenomena. It is an active field of study in linguistics or philosophy of language, but has received very little attention in computational linguistics.

6 Related work

Progovac (1995) begins with the radical syntactic change that Abney had recently proposed: that noun phrases are headed by the determiner, not the noun (Abney, 1987). She demonstrates, using the position of pronouns and other demonstratives, that Serbian-Croatian (a language without articles like *a/the*) has a determiner position, and that it is usually filled by a null determiner of some definiteness. In S-C, there is a definiteness marker *-i*, which Progovac explains as arising in determiner position, but transferring to the adjective on the surface.

Ko et al. (2009) uses L2-acquisition errors to demonstrate definiteness effects via determiner usage. They distinguish ‘definiteness’ and ‘specificity’ by stating that ‘definiteness’ involves common knowledge of uniqueness of referent between speaker and hearer, while ‘specificity’ only involves that uniqueness be known to the speaker. Native Korean and Russian speakers (both article-less languages) demonstrated, via a task very similar to the one described in this paper, that speakers tend to associate two articles, like *a* and *the*, with the value of one of these features, either [±definite] (correct, for English) or [±specific] (incorrect). Given some larger context and a placeholder in determiner position, the task is to replace the placeholder with either *the*, *a*, or a blank. While my task does not consider the null determiner option, this experimental design lends credibility to my peculiar and somewhat artificial task.

Similarly, Cho (1996) looks for evidence of a determiner phrase in Russian, in native English learners’ acquisition of Russian. Based on patterns of these speakers using certain definiteness-marking adjectives (with ‘covertly expressed’ definite features), she claims that the English grammar of definiteness is readily transferred to definiteness features in Russian.

Vieira and Poesio (2000) uses a portion of the Penn Treebank containing about $1,400$ definite descriptions, based on annotations gathered from multiple annotations produced by their subject pool (this was partly to determine how easy / reliable the task itself was, which they investigate further in Poesio and Vieira (1998)). They primarily had the purpose of distinguishing discourse-old and discourse-new descriptions; while they call their system shallow, it uses much more descriptive features than mine, using aspects such as copular constructions, specific predicates, like factives, restrictive / non-restrictive post-modification, appositivity, bridging features and and NER. Their prediction model is a decision tree based on these features, and their F-score on the test data set (400 of the $1,400$ descriptions) was 0.69.

6.1 Definiteness across translation

Siegel (1996) addresses the specific problem of Japanese to German machine translation, focusing on Japanese features that are not manifested at surface level. German determiners like *ein* and *dem* do not align to anything in the source Japanese sentences, but they are required in the German translation. Siegel derives Prolog transfer rules for inserting appropriate determiners in the German sentence output.

Siegel claims that this process requires more than just post-processing output in the target language, and that several factors affect the choice of determiners in the German. For her, the problem is 'interlingual,' and must be addressed at the level of translation. However, as most modern translation systems have shown, techniques ignoring theoretical complexities often surpass more linguistically correct systems.

Other approaches have included pre-processing Japanese text for relevant markers, inferring definiteness and number at the source language level. Her rules use other definiteness signals in the source language (Japanese), like numerals. Using a combination of preference rules with defaults, handles translating several types of Japanese noun phrases into German, inserting number and definiteness markers that are not given in the Japanese source sentences.

Conclusion

Siegel doesn't discuss translating German back to Japanese; is it sufficient to simply drop most German determiners and number morphology? If so, are there easy ways to predict which things get dropped? Are there other issues that arise in the interaction between these two languages? Further, I'm curious if Japanese–Russian translation is easier, because those languages both lack determiners.

My project is a cross-section of some of issues raised in Siegel (1996). Instead of looking at a particular language pair,

I'm investigating the recoverability of definiteness markers at the surface level—on the target language, in a translation scenario. This is more general; whereas language pairs are quadratic in the number of languages handled by the translation system, and translation between them requires a large number of parallel texts, surface-level post-processing depends only on the target language and does not require the same depth of annotation. This means much more data can be put to use, more easily.

7 Limitations and future work

Currently, this system uses POS tags provided with the corpus as features in the training as well as the testing phase. As previously mentioned, determiners are useful features in POS tagging, so this interdependence could potentially become a catch-22. However, I think that the difference between *a/an* and *the* should not be too egregious an issue for a POS tagger; I expect that a POS tagger trained on a corpus with anonymized determiners would learn to use the merged placeholder 'DET' just as well.

A bigger issue is that *a* and *the* are blindly used as indicators of definiteness. While the corpus developed in Vieira and Poesio (2000) is relatively small, their approach suggests that some measure of definiteness / specificity could be induced from crowdsourced annotations. Active learning could be used to label the less amibiguous determiners (like *the*, *that*), leaving definiteness annotations of more ambiguous determiners and quantifiers (like *a*, *this*, *some*) to human judges.

Appendix

Code written to perform the analyses in this project is available at github.com/chbrown/nlp.

References

Steven Abney. *The English noun phrase in its sentential aspect*. PhD thesis, MIT, Cambridge, Mass, 1987.

David Beaver. Centering and the optimization of discourse. http://goo.gl/UgyKk, July 2000.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. The NXT-format Switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Lang Resources & Evaluation*, 44:387–419, 2010. doi: 10.1007/s10579-010-9120-1.

Jacee Cho. Where is the feature [definite] encoded in Russian? : Empirical data from L2 acquisition. In *The Slavic Forum*, 1996. URL `http://lucian.uchicago.edu/blogs/theslavicforum/slavic-forum-2011`.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.*, 21:203–225, June 1995.

Kiyoshi Ishikawa. Crosslinguistic notions of (in)definiteness. In *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, 1995.

Heejeong Ko, Tania Ionin, and Ken Wexler. L2-acquisition of English articles by Korean speakers. *The handbook of East Asian psycholinguistics: Korean*, 3:286–304, 2009.

Massimo Poesio and Renata Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2), 1998.

Ljiljana Progovac. Determiner phrase in a language without determiners. *University of Venice: Working Papers in Linguistics*, 5(2), 1995.

Melanie Siegel. Preferences and defaults for definiteness and number in Japanese to German machine translation. In *Language, Information and Computation (PACLIC 11)*, pages 43–52, 1996.

Renata Vieira and Massimo Poesio. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593, 2000.