Inferring 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 definiteness markers, *a/an* 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.[1]

(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 indefinite; (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.

[1]Excluding the deictic sense of *this*.

Definiteness is generally defined as the form of noun phrase that is acceptable when the uniqueness of the 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; (4) is 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."[2]); *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.

I dont get these exceptions

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 cor-

[2]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).

pus 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 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 and 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 almost 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 starts relatively high; a maximum likelihood estimate using the more common of the two markers (*the*) produces an accuracy of 69.1%.

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).

need to first determine if article is needed at all to do generation.

I compare results of CRF (Conditional Random Field) sequence labeling against a SVM (Support Vector Machine) tagging each token alone. And, as this is primarily a feature selection problem, and I look at 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 (Brennan et al., 1987, Grosz et al., 1995, Beaver, 2000). When generating natural language from the logical form of a series of events,

describe method in a section first

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 level of indefiniteness, or are used for other special purposes (such as to denote humans). When translating from these languages to languages with overt definiteness markers, we must insert determiners that often have no parallel marker in the source text. There are a few papers discussing this particular problem in machine translation, but they are specific to only a few language pairs (Ishikawa, 1995, Siegel, 1996).

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." Some Russian adjectives mark definiteness, but otherwise, it must be inferred from context or left ambiguous. But translating from Russian into English requires choosing appropriate determiners, for which we need a model of definiteness (in conjunction with a model of anaphora or centering).

2.3 Other treatment

These applications are useful to two specific niches of NLP, and apply only in some cases of those areas, which may explain why so little computational work exists on the subject. The linguistic and philosophical literature on definiteness is much more developed, but most of the approaches from that area require deep understanding of language. This treatment is outside the scope of this paper, as I focus on shallow features.

3 Experiment

Because English has determiners, every document is 'labeled' data. However, because I want to use features deeper than simple surface tokens, I use the WSJ portion of the Penn Treebank (see Table 2).

2,038	articles
1,027,820	tokens
77,377	determiners

Table 2: WSJ overview.

To test definiteness inference, I simply replace *a(n)* and *the* with the placeholder «DET», while retaining the original token as the label for that token (collapsing *a* and *an* into one representation). While this is an artificial evaluation metric, it could be useful in both of the application scenarios.

In the language generation case, a noun phrase would surface with a placeholder inserted in determiner position, which could then be resolved by the approach in this paper, using nearby features to infer definiteness. Similarly, 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 this post-processing step performed on just the target language.

3.1 Features and Model

Here are the features that I use, along with the names describing them in the graphs that follow.

def na 'NA' if this particular token is not a «DET» position.

def token Either «DET» or the original token. Along with 'def na,' this feature is vacuous in the token-wise discrimination case, though useful for the CRF, which labels all tokens.

next def token Either «DET» or the original token for the position directly following the target position.

next noun The next token found in the sequence labeled with one of the Penn Treebank POS tags: NN, NNS, NNP, or NNPS.

not clear why using CRF since sequence info not seemingly relevant

Show examples of features, confusing

next noun seen Whether or not the 'next noun' occurs earlier in the article.

This is not a typical sequence labeling problem, since relatively few of the tokens need to be resolved. But each instance of definiteness is not independent of the previous—in fact, a prior mention is a strong indicator that an entity is now in the active context, and that 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.

Most of this sequential aspect of anaphora and shifting context is handled by the preprocessing step. However, I compare using an SVM[3] to using a CRF.[4] In addition to the learning method, I also compare feature sets, to determine which are the most efficient at predicting the deleted definiteness markers.

quite limited set of features, could motivate these better

but CRF just allows dependence on previous token

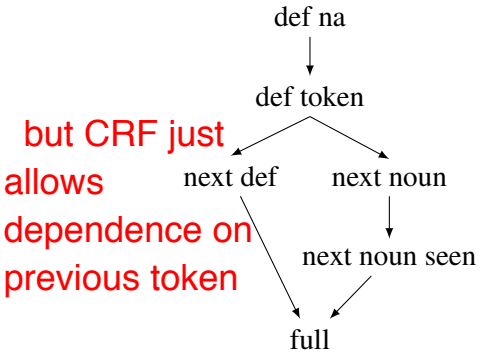

Figure 1: Hierarchy of inclusion of features in experimental feature sets.

3.2 Results

In each of the figures below, I have drawn the baseline as a horizontal line at 69.1% for simplicity; realistically, it would vary slightly across different cross-fold validation samples of the corpus.

I have sampled results for each of the following total counts of articles: 100, 250, 500, 1000, 2038 (the maximum). Each result reported is for a random shuffling of articles (in the CRF case) or tokens (in the SVM case), using a train / test split of 90% / 10%.

The legend in each plot describes different subsets of my feature functions. Most of these follow a pattern of incremental inclusion of features, see Figure 1 for the full hierarchy (each node includes all of its parents' features).

SVM results use a polynomial kernel. The first plot below shows that a polynomial kernel consistently performs better than alternative kernels; the accuracy lines depict results using the full feature set (see Figure 2).

The full feature set consistently performs the highest (see Figure 3). Differences between the next token and the next noun (which are often equivalent) are small; this is somewhat surprising, the next token

odd way to plot learning curve, use 10 fold CV

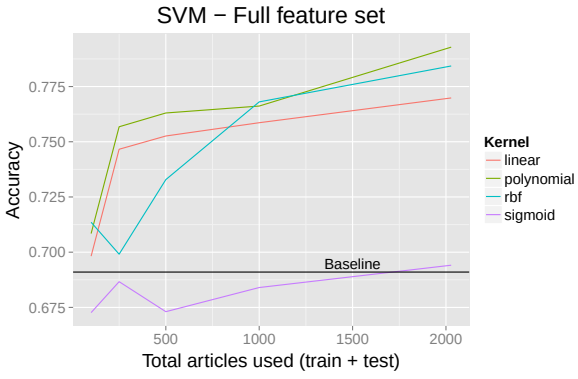

Figure 2: Comparison of SVM kernels.

[3] SVM^{light} (Joachims, 1999) via PySVMLight. I also tried LaSVM, but results were little better than the baseline.

[4] CRFSuite (Okazaki, 2007).

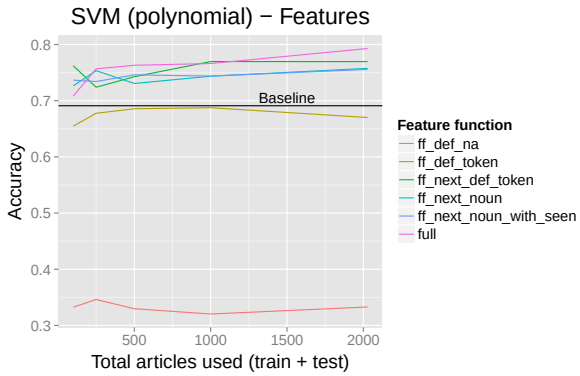

Figure 3: Comparison of feature sets for SVM.

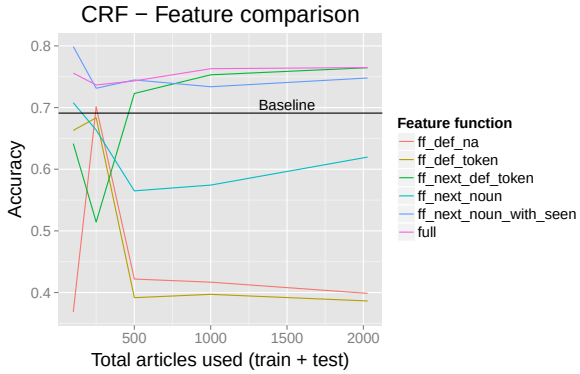

Figure 4: Comparison of feature sets for CRF.

Why?

is an easier metric but seems to do just as well or better. Searching the text of the article for the previous occurrences of the target noun does not seem to add any significant gain.

Discuss and analyze results

The CRF offered no benefit, though we see mostly the same pattern in trade-offs between more complex features sets (see Figure 4). The SVM seems better suited to the problem, and its results are a few percentage points higher, on average.

Notably, though, the CRF and the SVM ran in comparable time, despite the CRF also providing vacuous tags for all non-«DET» tokens.

These results demonstrate that it is relatively simple to increase accurate inference of definiteness on this pair of definite-marking articles.

4 Related work

Progovac (1995) begins with the radical syntactic change that Abney (1987) had recently proposed: that noun phrases are headed by the determiner, not the noun. 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 Serbian-Croatian, 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

Relation to your work?

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.

4.1 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.

5 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.

Better features!

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 crowd-sourced annotations. Active learning could be used to label the less ambiguous determiners (like *the*, *that*), leaving definiteness annotations of more ambiguous determiners and quantifiers (like *a*, *this*, *some*) to human judges.

6 Conclusion

For this task of collapsing *a/an* and *the* into a single token and then predicting the original identity, I have shown that an SVM (or CRF) can significantly improve upon the baseline by using a small and simple set of features. This approach is promising as part of the foundation of a larger language generation system, though such a system's requirements span far wider than the narrow issue developed here. However, I think that more linguistically justifiable approaches to problems can offer improvements to current solutions, even if the methodology is much more simple than its theoretical basis.

Implementing this, or some part of this approach, in a language generation system will be a much more accurate test of its worth, but I believe that any improvement in making generated language sound more natural is worthwhile. This paper shows that some aspect of definiteness is encoded in very simple features, which, if factored into the generation process, will help to determine what type of noun phrase is most appropriate.

Appendix

Code written to perform the analyses in this project is available at github.com/chbrown/nlp. See the `python/det` subfolder in particular.

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.

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, ACL '87, pages 155–162, Stroudsburg, PA, USA, 1987. Association for Computational Linguistics.

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 syn-

tax, 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.

Thorsten Joachims. Making large-scale svm learning practical. In B. Sch olkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

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.

Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (CRFs), 2007. URL `http://www.chokkan.org/software/crfsuite/`.

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.

Novel and interesting project topic and good general methodology, but details of method and evaluation are troubling and limited analysis and