Sense Anaphoric Pronouns: Am I *One*?

Marta Recasens
Google Research
1600 Amphitheatre Parkway
Mountain View, CA 94043
recasens@google.com

Zhichao Hu
Department of Computer Science
University of California, Santa Cruz
Santa Cruz, CA 95064
zhu@soe.ucsc.edu

Olivia Rhinehart
Google Research
1600 Amphitheatre Parkway
Mountain View, CA 94043
orhinehart@google.com

Abstract

This paper focuses on identity-of-sense anaphoric relations, in which the sense is shared but not the referent. We are not restricted to the pronoun *one*, the focus of the small body of previous NLP work on this phenomenon, but look at a wider range of pronouns (*that*, *some*, *another*, etc.). We develop annotation guidelines, enrich a third of English OntoNotes with sense anaphora annotations, and shed light onto this phenomenon from a corpus-based perspective. We release the annotated data as part of the SAnaNotes corpus. We also use this corpus to develop a learning-based classifier to identify sense anaphoric uses, showing both the power and limitations of local features.

1 Introduction

Anaphora and coreference are two linguistic phenomena that often occur together and as a result are sometimes regarded to be the same, but one can happen without the other. Compare (1) and (2)[1]: in the former *them* is anaphoric since its interpretation depends upon another expression in the context (*multiple loans*) with which it also corefers; in the latter example, in contrast, *ones* anaphorically depends on *loans* but they do not corefer since the demanded and existing loans are different discourse entities.

(1) If you have *multiple loans*, you can consolidate **them** into a single loan.

(2) Consumers and companies demand fewer *loans* and struggle to pay back existing **ones**.

[1]Anaphors are shown in bold, and antecedents in italics.

In this paper we focus on (2)-like anaphoric relations, which have been called *identity-of-sense anaphora* (Grinder and Postal, 1971; Hirst, 1981)—we use the term *sense anaphora* for short. Sense anaphoric pronouns inherit the sense from their antecedent but do not denote the same referent. The meaning of these anaphoric forms remains empty if they are not identified and resolved, thus the importance for most natural language understanding tasks.

While a good deal of previous work in linguistics (Grinder and Postal, 1971; Bresnan, 1971) has discussed the underlying syntactic representation of this phenomenon (syntactic deletion, interpretive rules, etc.) and closely related ones like noun ellipsis (Sag, 1976; Dalrymple et al., 1991), there have been few computational studies on sense anaphora and these have focused on the pronoun *one* (Gardiner, 2003; Ng et al., 2005).

We target a wider set of sense anaphoric pronouns, not only *one* but also *that*, *many*, *some* and others. We annotate a third of English OntoNotes (Weischedel et al., 2011) with sense anaphoric pronouns together with their antecedent, going beyond the existing coreference annotations. To our knowledge, this is the first annotation effort of this phenomenon on a large corpus, and it uncovers distributional statistics and real-world usage patterns.

Our second contribution is using the annotated data to train a learning-based system that identifies anaphoric uses (2) from non-anaphoric uses of the target pronouns such as generics (3), when their meaning is equivalent to 'some people'.

(3) While **some** think that the estimate may be inflated, the consensus is that drier seasons are on the horizon.

By using only local lexical and syntactic features, the system reaches 79.01% F1 on *one*, and 67.34% F1 on an extended list of sense anaphors.

2 Sense Anaphora

The majority of studies on identity-of-sense anaphora (Grinder and Postal, 1971; Bresnan, 1971) focus on verb rather than noun phrases. We focus on the latter, considering not only *one*, but also other expressions that can similarly borrow their sense from a contextual expression (4), (6), (8). We target single-token anaphors in the following categories:

- ONE (*one, ones*)
- Quantifiers
  - INDEFINITES (*all, any, few, many, more, most, much, some*)
  - NUMERALS (*two, three, ..., hundred, etc.*)
  - MEASURE NOUNS (*bit(s), bunch, couple, dozen(s), lot(s), pair, plenty, ton(s)*)
- DEMONSTRATIVES (*that, those*)
- POSSESSIVES (*mine, yours, his, hers, ours, theirs*)
- OTHER (*other(s), another*)

Previous studies (Dahl, 1985; Luperfoy, 1991; Gardiner, 2003; Payne et al., 2013) have classified *one* in terms of its uses—determiner, count noun, pronoun, etc.—and antecedent types—a kind, a set, an individual. Drawing on these previous classifications, we extend them to all types of sense anaphora and consider that every (sense) anaphor–antecedent pairing falls into one of two broad classes:

Partitive Denotes a subset relationship between the anaphor and antecedent (the set), where the anaphor not only shares the sense with the antecedent but also the specified characteristics. For this reason, the whole noun phrase is considered the antecedent (4). Non-anaphoric partitives, those followed by *of* plus the set (5), are excluded, following Gardiner (2003).[2]

(4) The blast kills *two cameramen*, **one** from Spanish TV, **another** from Reuters.

(5) That's **one** of the problems that they are facing so far.

[2]Note that Gardiner (2003) uses the term *partitive* exclusively for non-anaphoric partitive uses of *one*, i.e., *one of*.

The partitive class is similar to bridging anaphors of the set-membership or subset types (Clark, 1975; Poesio et al., 1999; Markert et al., 2012), but we focus on the pronoun-like sense anaphors listed above (which can be seen as headless noun phrases), whereas bridging anaphors usually target full noun phrases, e.g., *a group of students . . . three boys*.

Instantiator The anaphor is a new instance created from the same sense as the antecedent (6), (7).

(6) In both quantity and quality, the *English teaching materials* of today leave **those** of before in the dust.

(7) High-tech *industries* need to be constantly innovative, while traditional **ones** have to undergo transformation.

The newly created instance may inherit only the core sense or include some of the specifics of the antecedent, so antecedent spans only include the inherited modifiers: all of them in the case of (6), or none in the case of (7), where inheriting the modifier *high-tech* would contradict the anaphor's modifier, *traditional*. The category of *other*-anaphora or comparative anaphora (Markert and Nissim, 2005) falls into the instantiator class, but similarly to set-membership bridging, the study of *other*-anaphora has focused on full noun phrases rather than headless ones.

The line between the partitive and instantiator classes can be blurry, especially when the antecedent is a kind rather than a set (8), but we find the distinction helpful to conceptualize sense anaphora.

(8) He has done research on *traditional Chinese poetry*, and has included **some** in his Portuguese-language writings.

3 SAnaNotes

3.1 Annotation

Annotators identified sense anaphors and their antecedent phrases using a custom GUI. Strings from the target anaphor categories were automatically highlighted in the text; annotators first determined if the highlighted tokens were anaphors and, if so, they identified their corresponding antecedent. Considering whether the pairing belonged to the partitive or instantiator class helped them determine what the boundaries of the antecedent were, but the partitive/instantiator distinction was not annotated.

When the antecedent is part of a coreference chain, annotators chose the phrase that directly preceded the anaphor, unless it was a relative pronoun. In (9), *it* is chosen as the antecedent instead of *debate* because it is the closest mention in the antecedent's coreference chain. The tool allowed for two anaphors to share the same antecedent (10).

(9) That debate, *it*'s a hard **one** when Hardball returns.

(10) Added to this is the perennial problem of class sizes being too large, and not enough *English classes* scheduled – only **one** or **two** a week.

There was also a 'no explicit antecedent' option for cases in which an anaphor borrows its sense from an antecedent that is not available in the text. In (11), both anaphors *ones* inherit a sort-of-*issue* sense, but this antecedent is not explicit but built up in the context of the passage.

(11) It must be advanced with a plan, the easy **ones** first and the tough **ones** last.

Four human annotators participated in the annotation. After an initial pilot training period, they completed single-annotation on 1 138 documents. In a final stage of annotation, the annotators completed four-way annotation on a set of 25 documents to measure inter-annotator agreement. We used Fleiss' (1971) kappa to measure agreement on anaphor identification: $\kappa = .67$, which indicates substantial agreement according to Landis and Koch (1977). For the commonly identified anaphors, pairwise agreement on their antecedent spans was 63%. The most common annotation errors are anaphor omissions: given the small percentage of anaphoric uses for some of the categories (see Table 1), sense anaphors are easy to be missed.

3.2 Data

The source data for annotation was a third of the English documents from OntoNotes (Weischedel et al., 2011), a 1.6-million-word corpus covering a variety of domains (newswire, broadcast conversation, weblogs, magazine, New Testament, etc.), sampled so as to keep the proportion of OntoNotes domains. We annotated 1 163 documents in total. We release this annotated corpus as SAnaNotes, available from `https://github.com/dmorr-google/sense-anaphora`.

Token	TRAIN			TEST		
	Freq.	Ana.	%	Freq.	Ana.	%
one	1 099	148	13.5	268	33	12.3
ones	49	29	59.2	8	3	37.5
all	720	13	1.8	185	0	0.0
another	41	28	68.3	13	8	61.5
few	142	11	7.7	45	3	6.7
many	448	18	4.0	121	3	2.5
more	846	13	1.5	231	0	0.0
most	336	6	1.8	103	1	1.0
much	369	1	0.3	90	1	1.1
other	597	21	3.5	141	5	3.5
others	139	43	30.9	32	11	34.4
some	214	26	12.1	49	5	10.2
that	940	25	2.7	248	13	5.2
those	296	27	9.1	65	13	20.0
NUM	6 046	120	2.0	1 880	16	0.9
TOTAL	12 282	529	4.3	3 479	115	3.3

Table 1: Distribution of sense anaphors in SAnaNotes (*Ana.* stands for 'anaphoric'). TRAIN subsumes the development data. The Freq. column excludes determiner uses.

The average number of sense anaphors per document is 0.6. Of the target categories, the OntoNotes data contain a small number of POSSESSIVES (*hers, yours*, etc.) and MEASURE NOUNS (*bunch, ton*, etc.), of which anaphoric examples represent an even smaller number. Table 1 shows the distribution of anaphors belonging to categories for which there are at least 10 anaphoric examples, that is, keeping ONE, INDEFINITES, NUMERALS, DEMONSTRATIVES, and OTHER; and excluding POSSESSIVES and MEASURE NOUNS. While the ONE and OTHER classes show a large proportion of anaphoric uses, and DEMONSTRATIVES to a smaller extent, only a small number of INDEFINITES and NUMERALS are anaphoric.

4 Anaphoric Classification

Using the SAnaNotes corpus, we built a classifier to distinguish sense anaphors (example 2) from other uses like determiners, numerals, generics (example 3), deictics, etc. Given the composition of SAnaNotes, we target all anaphors listed in Table 1.

4.1 Previous Work

To our knowledge the only computational approaches to resolving sense anaphoric pronouns have focused on *one* anaphora, namely the stud-

ies by Gardiner (2003) and Ng et al. (2005). Both split the problem into two steps: identification of anaphoric uses and resolution to an antecedent. We overview the first step since it is our current focus.

Gardiner (2003) developed a rule-based system based on five heuristics to distinguish non-anaphoric uses—numeric (*one* is a quantifier or numeric adjective), partitive (*one*'s immediate post-modifier is *of* introducing a plural noun phrase), generic (*one* is a subject of a modal or animate verb)—from anaphoric ones (the rest). She extracted from the British National Corpus a test set of 773 sentences containing *one*, but highly biased towards anaphoric examples (71.5%) and far from reality (compare with 12.3% in Table 1). On this test set her system obtained 85.4% precision and 86.9% recall.

Ng et al. (2005) developed a learning-based system by turning Gardiner's (2003) heuristics into seven learning features. They trained a C4.5 decision tree classifier using 10-fold cross validation on a set of 1 577 *one* expressions, also from the British National Corpus, but this time randomly selected, thus mirroring the natural distribution of anaphoric *one* (15.2% in their data set). They obtained 68.3% precision and 80.0% recall, and noted that discriminating between the anaphoric and generic classes offered the most complexity out of all six classes.

4.2 System

In contrast to previous work, our goal is to address a wider variety of sense anaphors and to use simple lexical and syntactic features that could identify the constructions characteristic of sense anaphoric uses, e.g., anaphoric *that* is usually followed by an *of*-phrase, generic *one* is often the subject of specific animate verbs, etc.

We generate a training instance for every token matching one of those in Table 1 and every token with NUM category, and exclude determiners (tokens with 'det' label). Filtering 'num' or 'amod' labels gave poorer results on the development set and so we kept them, leaving it to the classifier to learn when to filter them out. Given the multiple senses of *that*, we exclude its uses as a relative pronoun (tag: 'WDT') or conjunction (tag: 'IN').

We train an SVM classifier–LIBLINEAR implementation (Fan et al., 2008)–to distinguish between the anaphoric class and the rest using 31 lexical and syntactic feature types:

- Lowercased word, POS tag, dependency label and word cluster from Brown et al. (1992) for:
  - Anaphor candidate
  - Two previous tokens
  - Two following tokens
  - Candidate's syntactic head (e.g., *says* is the head of *another* in *another says* . . .)
  - Candidate's syntactic leftmost child (e.g., *the* is the leftmost dependent of *one* in *the second one he has missed*)
- Conjoined features with these pairs:
  - Lowercased candidate and POS tag of the previous token (and vice versa)
  - Lowercased candidate and POS tag of the following token (and vice versa)
  - Lowercased candidate and leftmost child

We also try adding the finer-grained features used by Ng et al. (2005), but they do not help, presumably because they are already covered.

4.3 Evaluation

We split the SAnaNotes corpus into train, development, and test partitions. Once development was over, we merged that partition with the train set. The anaphoric class usually represents a small percentage of all occurrences of every candidate token (Table 1). For the feature generation, we annotated the data with a dependency parser similar to MaltParser (Nivre et al., 2007).

We use precision (P) and recall (R) to measure the number of correct anaphoric predictions made by our system, and the proportion of gold anaphors identified by our system.

For comparison, we reimplemented the two previous systems that focused on *one*: (1) the rule-based system by Gardiner (2003), and (2) an unpruned J48 decision tree classifier trained with Ng et al.'s (2005) features in Weka 3.6.12 (Hall et al., 2009), using the same train/test SAnaNotes split.[3]

4.4 Results and Discussion

Table 2 compares the results of our system on *one* and *ones* with those obtained by the two previous

[3]The scores for a pruned decision tree were all 0.

System	P	R	F1
Gardiner (2003)	40.00	94.44	56.20
Ng et al. (2005)	62.50	55.60	58.80
Our system	74.42	88.89	79.01

Table 2: Comparison of our *one*-anaphor classifier (including *one* and *ones*) with previous work on the test set of SAnaNotes.

Anaphor class	P	R	F1
ONE	71.11	88.89	79.01
INDEFINITES	38.46	38.46	38.46
NUMERALS	27.78	31.25	29.41
DEMONSTRATIVES	87.50	53.85	66.67
OTHER	61.54	66.67	64.00
ALL	61.02	62.61	61.80
ALL excl. NUMERALS	67.00	67.68	67.34

Table 3: Evaluation of our anaphor classifier on the test set of SAnaNotes.

systems on the same test set from SAnaNotes. Gardiner's precision is especially low, which did not show in her original test set highly biased towards anaphoric instances. Our features, though less targeted to the ones used by Ng et al. (2005), turn out to perform better. In addition, they generalize to the additional sense anaphors not tackled before. Ng et al.'s features are very specific and fail to learn to discriminate between some anaphoric patterns that our more general system learns (12).

(12) The hottest gift this Christmas could be *Sony's new PlayStation 2*, but good luck finding **one**.

Table 3 breaks down the performance of our anaphor classification results by anaphor category. Classification of anaphors other than *one* and *ones* is considerably harder, especially for numerals, followed by indefinites. The small number of positive training instances (Table 1) probably accounts for the poor performance. Given the limited number of sense anaphors per document, a larger corpus is likely to make a significant difference. Our feature set generalizes better than Ng et al.'s (2005), which only obtains 38.60% F1 on all anaphors excluding numerals (vs. 67.34% F1 by our system) and 31.30% F1 when including numerals (vs. 61.80% F1 by our system).

The classification errors illustrate the complexity of the task: (13) is a precision error given that *few* refers to the number itself, but it is arguably a borderline case; (14) is a recall error that shows the limitation of surface features in a small context window because *some* would be interpreted generically if there was no previous context.

(13) They were able to whittle it down the number of missing aircraft uh to a **few**.

(14) *Today's Tanshui residents* are living their own stories [...] **Some** are active in the morning, **some** late at night.

5 Conclusion

We tackled a tail phenomenon in natural language understanding, that of sense anaphora, going beyond the pronoun *one* and generalizing to other similar identity-of-sense expressions. While not very common in the OntoNotes domains, we suspect they are more common in conversational language, for example voice queries, thus being especially important for the next generation of voice assistants.

Apart from annotating and releasing SAnaNotes, we experimented with the anaphoric classification task, achieving 61.80% F1 with a set of local features. As future work, we would like to approach the antecedent resolution task jointly with anaphor identification, as the hardest cases of the anaphoric/generic distinction require knowing whether an antecedent is available in the context. This would also make it possible to explore features that look at a wider context, for example to capture parallel structures between antecedent and anaphor (e.g., *high-tech industries* and *traditional ones*) as well as features that take discourse structure into account, e.g., discourse relations such as comparison and conjunction between the discourse units containing the antecedent and anaphor.

Acknowledgments

This work was done when the second author was an intern at Google. We would like to thank all the annotators who worked on SAnaNotes (Melanie Bolla, Katy DiNatale, Grace Gaspardo, Patrick Hegarty) as well as Amanda Morris and Edgar Gonzàlez for their support and productive discussions, and Dave Orr for making the data release possible. Many thanks also to the nNLP team, especially Qiwei Wang and David Huynh, for customizing the annotation tool for this task. Finally, we would like to thank our anonymous reviewers for their helpful comments.

References

Joan Bresnan. 1971. A note on the notion "identity of sense anaphora". *Linguistic Inquiry*, 2:589–597.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Herbert H. Clark. 1975. Bridging. In R. C. Schank and B. L. Nash-Webber, editors, *Theoretical Issues in Natural Language Processing*, pages 169–174.

Deborah Anna Dahl. 1985. *The structure and function of one-anaphora in English*. Ph.D. thesis, University of Minnesota.

Mary Dalrymple, Stuart M. Shieber, and Fernando C.N. Pereira. 1991. Ellipsis and higher order unification. *Linguistics and Philosophy*, 14:399–452.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Mary Gardiner. 2003. Identifying and resolving *one*-anaphora. B.S. Thesis, Department of Computing, Macquarie University.

John Grinder and Paul M. Postal. 1971. Missing antecedents. *Linguistic Inquiry*, 2:589–597.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11:10–18.

Grame Hirst. 1981. *Anaphora in natural language understanding: a survey*. Lecture Notes in Computer Science. Springer-Verlag, Berlin, New York.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Susann Luperfoy. 1991. *Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions*. Ph.D. thesis, University of Texas at Austin.

Katja Markert and Malvina Nissim. 2005. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31:367–401.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of ACL*, pages 795–804.

Hwee Tou Ng, Yu Zhou, Robert Dale, and Mary Gardiner. 2005. A machine learning approach to identification and resolution of *one*-anaphora. In *Proceedings of IJCAI*, pages 1105–1110.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13:95–135.

John Payne, Geoffrey K. Pullum, Barbara C. Scholz, and Eva Berlage. 2013. Anaphoric *one* and its implications. *Language*, 89(4):794–829.

Massimo Poesio, Florence Bruneseaux, and Laurent Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Proceedings of the ACL Workshop Towards Standards and Tools for Discourse Tagging*, pages 65–74.

Ivan A. Sag. 1976. *Deletion and logical form*. Ph.D. thesis, Massachusetts Institute of Technology.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 54–63. Springer.