Antecedent Selection for Sluicing: Structure and Content

Pranav Anand
Linguistics
UC Santa Cruz
panand@ucsc.edu

Daniel Hardt
IT Management
Copenhagen Business School
dh.itm@cbs.dk

Abstract

Sluicing is an elliptical process where the majority of a question can go unpronounced as long as there is a salient antecedent in previous discourse. This paper considers the task of antecedent selection: finding the correct antecedent for a given case of sluicing. We argue that both syntactic and discourse relationships are important in antecedent selection, and we construct linguistically sophisticated features that describe the relevant relationships. We also define features that describe the relation of the content of the antecedent and the sluice type. We develop a linear model which achieves accuracy of 72.4%, a substantial improvement over a strong manually constructed baseline. Feature analysis confirms that both syntactic and discourse features are important in antecedent selection.

1 Introduction

Ellipsis involves sentences with missing subparts, where those subparts must be interpretatively filled in by the hearer. How this is possible has been a major topic in linguistic theory for decades (Sag, 1976; Chung et al., 1995; Merchant, 2001). One widely studied example is *verb phrase ellipsis* (VPE), exemplified by (1).

(1) Harry traveled to southern Denmark to study botany . Tom did too .

In the second sentence (*Tom did too*) the verb phrase is entirely missing, yet the hearer effortlessly 're-solves' (understands) its content to be *traveled to southern Denmark to study botany*.

Another widely studied case of ellipsis is *sluicing*, in which the majority of a question is unpronounced, as in (2).

(2) Harry traveled to southern Denmark to study botany . I want to know **why** .

Here the content of the question, introduced by the WH-phrase *why*, is missing, yet it is understood by the hearer to be *why did Harry travel to southern Denmark to study botany?*. In both of these cases, ellipsis resolution is made possible by the presence of an *antecedent*, material in prior discourse that, informally speaking, is equivalent to what is missing.

Ellipsis poses an important challenge for many applications in language technology, as various forms of ellipsis are known to be frequent in a variety of languages and text types. This is perhaps most evident in the case of question-answering systems, since elliptical questions and elliptical answers are both very common in discourse. A computational system that can effectively deal with ellipsis involves three subtasks (Nielsen, 2005): *ellipsis detection*, in which a case of ellipsis is identified, *antecedent selection*, in which the antecedent for a case of ellipsis is found, and *ellipsis resolution*, where the content of the ellipsis is filled in with reference to the antecedent and the context of the ellipsis. Here, we focus on antecedent selection for sluicing. In addressing this problem of antecedent selection, we make use of a newly available annotated corpus of sluice occurrences (Anand and McCloskey, 2015). This corpus consists of 4100 automatically parsed and annotated examples from the New York Times subset of the Gigaword Corpus, of which 2185 are

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1234–1243,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

publicly available.

Sluicing antecedent selection might appear simple – after all, it typically involves a sentential expression in the nearby context. However, analysis of the annotated corpus data reveals surprising ambiguity in the identification of the antecedent for sluicing.

In what follows, we describe a series of algorithms and models for antecedent selection in sluicing. Following section 2 on background, we describe our dataset in section 3. Then in section 4, we describe the structural factors that we have identified as relevant for antecedent selection. In section 5, we look at ways in which the content of the sluice and the content of the antecedent tend to be related to each other: we address lexical overlap, as well as the probabilistic relation of head verbs to WH-phrase types, and the relation of correlate expressions to sluice types. In section 6 we present two manually constructed baseline classifiers, and then we describe an approach to automatically tuning weights for the complete set of features. In section 7 we present the results of these algorithms and models, including results involving various subsets of features, to better understand their contributions to the overall results. Finally in section 8 we discuss the results in light of plans for future work.

2 Background

2.1 Sluicing and ellipsis

Sluicing is formally defined in theoretical linguistics as ellipsis of a question, leaving only a WH-phrase *remnant*. While VPE is licensed only by a small series of auxiliaries (e.g., modals, *do*, see Lobeck (1995)), sluicing can occur wherever questions can, both in unembedded 'root' environments (e.g., *Why?*) or governed by the range of expressions that embed questions, like *know* in (2). Sluicing is argued to be possible principally in contexts where there is uncertainty or vagueness about an issue (Ginzburg and Sag, 2000). In some cases, this manifests as a *correlate*, an overt indefinite expression whose value is not further specified, like *one of the candidates* in (3). But in many others, like that in (2) or (4), there is no correlate, and the uncertainty is implicit.

(3) They 've made an offer to [$_{cor}$ one of the candidates] , but I 'm not sure **which one**

(4) They were firing , but **at what** was unclear

The existence of correlate-sluices suggests an obvious potential feature type for antecedent detection. However, the annotated sluices in (Anand and McCloskey, 2015) have correlates only 22% of the time, making this process considerably harder. We return to the question of correlates in section 5.1.

2.2 Related Work

The first large-scale study of ellipsis is due to Hardt (1997), which addresses VPE. Examining 644 cases of VPE in the Penn Treebank, Hardt presents a manually constructed algorithm for locating the antecedent for VPE, and reports accuracy of 75% to 94.8%, depending on whether the metric used requires exact match or more liberal overlap or containment. Several preference factors for choosing VPE antecedents are identified (Recency, Clausal Relations, Parallelism, and Quotation). One of the central components of the analysis is the identification of structural constraints which rule out antecedents that improperly contain the ellipsis site, an issue we also address here for sluicing. Drawing on 1510 instances of VPE in both the British National Corpus (BNC) and the Penn Treebank, Nielsen (2005) shows that a maxent classifier using refinements of Hardt's features can achieve roughly similar results to Hardt's, but that additional lexical features do not help appreciably.

Nielsen chooses to optimize for Hardt's Head Overlap metric, which assigns success to any candidate containing/contained in the correct antecedent. There are thus many "correct" antecedents for a given instance of VPE, which mitigates the class imbalance problem. However, the approach does not provide a way to discriminate between these containing candidates, an important step in the eventual goal of resolving the ellipsis.

There is no similar work on antecedent selection for sluicing, though there have been small-scale corpora gathered for sluices (Nykiel, 2010; Beecher, 2008). In addition, Fernandez et al. (2005) build rule-based and memory-based classifiers for the pragmatic import of root (unembedded) sluices in the BNC, based on the typology of Ginzburg and Sag (2000). Using features for the type of WH-phrase, markers of mood (declarative/interrogative) and polarity (positive/negative) as well as the presence of correlate-like material (e.g., quantifiers, definites, etc.), they can diagnose the purpose of a sluice in a dataset of 300 root sluices with 79% average F-score, a 5% improvement over the MLE. Fernandez et al. (2007) address the problem of identifying sluices and other non-sentential utterances. We don't address that problem in the current work. Furthermore, Fernandez et al. (2007) and Fernandez et al. (2008) address the general problem of non-sentential utterances or fragments in dialogue, including sluices. Sluicing in dialogue differs from sluicing in written text in various ways: there is a high proportion of root sluices, and antecedent selection is likely mitigated by the length of utterances and the order of conversation. As we discuss, many of our newswire sluices evince difficult patterns of containment inside the antecedent (particularly what we call interpolated and cataphoric sluices), and it does not appear from inspection that root sluices ever participate in such processes.

Looking more generally, there is an obvious potential connection between antecedent selection for ellipsis and the problem of coreference resolution (see Hardt (1999) for an explicit theoretical link between the two). However, entity coreference resolution is a problem with two major differences from ellipsis antecedent detection: a) the antecedent and anaphor often share a variety of syntactic, semantic, and morphological characteristics that can be featurally exploited; b) entity expressions in a text are often densely coreferent, which can help provide proxies for discourse salience of an entity.

In contrast, abstract anaphora, particularly discourse anaphora (*this/that* anaphora to something sentential), may offer a more parallel case to ours. Here, Kolhatkar et al. (2013) use a combination of syntactic type, syntactic/word context, length, and lexical features to identify the antecedents of anaphoric shell nouns (*this fact*) with precision from 0.35-0.72. Because of the sparsity of these cases, Kolhatkar et al. use Denis and Baldridge's (2008) candidate ranking model (versus a standard mention-pair model (Soon et al., 2001)), in which all potential candidates for an anaphor receive a relative rank in the overall candidate pool. In this paper, we will pursue a hillclimbing approach to antecedent selection, inspired by the candidate ranking scheme.

3 Data

3.1 The Annotated Dataset

Our dataset, described in Anand and McCloskey (2015), consists of 4100 sluicing examples from the New York Times subset of the Gigaword Corpus, 2nd edition. This dataset is the first systematic, exhaustive corpus of sluicing.[1] Each example is annotated with four main tags, given in terms of token sequence offsets: the *sluice remnant*, the *antecedent*, and then inside the antecedent the main *predicate* and the *correlate*, if any. The annotations also provide a free-text resolution. Of the 4100 annotated, 2185 sluices have been made publicly available; we use that smaller dataset here. We make use of the annotation of the antecedent and remnant tags. See Anand and McCloskey (2015) for additional information on the dataset and the annotation scheme. For the feature extraction in section 4, we rely on the the token, parsetree, and dependency parse information in Annotated Gigaword (extracted from Stanford CoreNLP).

3.2 Defining the Correct Antecedent

Because of disagreements with the automatic parses of their data, Anand and McCloskey (2015) had annotators tag token sequences, not parsetree constituents. As a result, 10% of the annotations are not sentence-level (i.e., S, SBAR, SBARQ) constituents, such as the VP antecedent in (5), and 15% are not constituents at all, such as the case of (6), where the parse lacks an S node excluding the initial temporal clause. We describe two different ways to define what will count as the correct antecedent in building and assessing our models.

3.2.1 Constituent-Based Accuracy

Linguists generally agree that the antecedent for sluicing is a sentential constituent (see Merchant (2001) and references therein). Thus, it is straightforward to define the antecedent as the minimal

[1] 4100 sluices works out to roughly 0.14% of WH-phrases in the NYT portion of Gigaword. However, note that this includes all uses of WH-phrases (e.g., clefts and relative clauses), whereas sluicing is only possible for WH-questions. It's not clear how many questions there are in the dataset (distinguishing questions and other WH-phrases is non-trivial).

sentence-level constituent containing the token sequence marked as the antecedent. Then we define CONACCURACY as the percentage of cases in which the system selects the correct antecedent, as defined here.

While it is linguistically appealing to uniformly define candidates as sentential constituents, the annotator choices are sometimes not parsed that way, as in the following examples:

(5) “ I do n’t know how , ” said Mrs. Kitayeva , “ but [$_S$ we want [$_{VP}$ **to bring Lydia home**] , in any condition] . ”

(6) [$_S$ [$_S BAR$ When Brown , an all-America tight end , was selected in the first round in 1992] **he was one of the highest rated players on the Giants ’ draft board**]

In such cases, there is a risk that we will not accurately assess the performance of our systems, since the system choice and annotator choice will only partially overlap.

3.2.2 Token-Based Precision and Recall

Here we define a metric which calculates the precision and recall of individual token occurrences, following Bos and Spenader (2011) (see also Kolhatkar and Hirst (2012)). This will accurately reflect the discrepancy in examples like (5) – according to ConAccuracy, a system choice of *we want to bring Lydia home in any condition* is simply considered correct, as it is the smallest sentential consituent containing the annotator choice. According to the Token-Based metric, we see that the system achieves recall of 1; however, since the system includes six extraneous tokens, precision is .4. We define TOKF as the harmonic mean of Token-Based Precision and Recall; for (5), TokF is .57.

3.3 Development and Test Data

The dataset consists of 2185 sluices extracted from the New York Times between July 1994 and December 2000. For feature development, we segmented the data into a development set (DS) of the 453 sluices from July 1994 to December 1995. The experiments in section 6 were carried out on a test set (TS) of the 1732 sluices in the remainder of the dataset, January 1996 to December 2000.

4 Structure

Under our assumptions, the candidate antecedent set for a given sluice is the set of all sentence-level parsetree constituents within a n-sentence radius around the sluice sentence (based on DS, we set $n = 2$). Because sentence-level constituents embed, in DS there are on average 6.4 candidate antecedents per sluice. However, because ellipsis resolution involves identification of an antecedent, we assume that it, like anaphora resolution, should be sensitive to the overall salience of the antecedent. This means that there should be, in principle, proxies for salience that we can exploit to diagnose the plausibility of a candidate for sluicing in general. We consider four principle kinds of proxies: measures of candidate-sluice distance, measures of candidate-sluice containment, measures of candidate ‘main point’, and candidate-sluice discourse relation markers.

4.1 Distance

Within DS, 63% of antecedents are within the same sentence as the sluice site, and 33% are in the immediately preceding sentence. In terms of candidates, the antecedent is on average the 5th candidate from the end of the n-sentence window. The positive integer-valued feature DISTANCE tracks these notions of recency, where DISTANCE is 1 if the candidate is the candidate immediately preceding or following the sluice site (DISTANCE is defined to be 0 only for infinitival Ss like S0 in (7) below). The feature FOLLOWS marks whether a candidate follows the sluice.

4.2 Containment

As two-thirds of the antecedents are in the same sentence as the sluice, we need measures to distinguish the candidates internal to the sentence containing the sluice. In general, we want to exclude any candidate that ‘contains’ (i.e., dominates) the sluice, such as S0 and S-1 in (7). One might have thought that we want to always exclude the entire sentence (here, S-4) as well, but there are several cases where the smallest sentence-level constituent containing the annotated antecedent dominates the sluice, including: parenthetical sluices inside the antecedent (8), sluices in subordinating clauses (9), or sluice VPs coordinated with the antecedent VP (10). We thus need features to mark when such candidates are 'non-containers'.

(7) [$_{S-4}$ [$_{S-3}$ I have concluded that [$_{S-2}$ I can not support the nomination] , and [$_{S-1}$ I need [$_{S0}$ to explain **why**]].]

(8) [$_{S-2}$ A major part of the increase in coverage , [$_{S-1}$ though Mitchell 's aides could not say just **how much** ,] would come from a provision providing insurance for children and pregnant women .]

(9) [$_{S-3}$ Weltlich still plans [$_{S-2}$ to go , [$_{S-1}$ although he does n't know **where**]]]

(10) [$_{S-2}$ State regulators have ordered 20th Century Industries Inc. [$_{S-1}$ to begin paying $ 119 million in Proposition 103 rebates or explain **why not** by Nov. 14 .]]

Conceptually, what renders S-3 in (9), S-2 in (8), and S-1 in (10) non-containers is that in all three cases the sluice is semantically dissociable from the rest of the sentence. We provide three features to mark this. First, the boolean feature SLUICEINPARENTHETICAL marks when the sluice is dominated by a parenthetical (a PRN node in the parse or an *(al)though* SBAR delimited by punctuation). Second, SLUICEINCOORDVP marks the configuration exemplified (10).

We also compute a less structure-specific measure of whether the candidate is meaningful once the sluice (and material dependent on it) is removed. This means determining, for example, that S-4 in (7) is meaningful once *to explain why .* is removed but S-1 is not. But the latter result follows from the fact that the main predicate of S-1, *need* takes the sluice govering verb *explain* as an argument, and hence removing that argument renders it semantically incomplete. We operationalize this in terms of complement dependency relations. We first locate the largest subgraph containing the sluice in a chain of *ccomp* and *xcomp* relations. This gives us gov_{max}, the highest such governor (i.e., *explain*) in Fig. 1. The subgraph dependent on gov_{max} is then removed, as indicated by the grayed boxes in Fig 1. If the resulting subgraph contains a verbal governor, the candidate is meaningful and CONTAINSSLUICE is false. By this logic, S-4 in (7) is meaningful because it contains *concluded*, but S-1 is not, because there is no verbal material remaining.

4.3 Discourse Structure

It has often been suggested (Asher, 1993; Hardt, 1997; Hardt and Romero, 2004) that the antecedent selection process is very closely tied to discourse relations, in the sense that there is a strong preference or even a requirement for a discourse relation between the antecedent and ellipsis.

Here we define several features that indicate either that a discourse relation is present or is not present.

We begin with features indicating that a discourse relation is not present: the theoretical linguistics literature on sluicing has noted that antecedents not in the 'main point' of an assertion (e.g., ones in appositives (AnderBois, 2014) or relative clauses (Cantor, 2013)) are very poor antecedents for sluices, presumably because their content is not very salient. The boolean features CANDINPARENTHETICAL (determined as for the sluice above) and CANDINRELCLAUSE mark these patterns.[2]

We also define features that would tend to indicate the presence of a discourse relation. These have to do with antecedents that occur after the sluice. Although antecedents overwhelmingly occur prior to sluices, we observe one prominent cataphoric pattern in DS, where the sentence containing the sluice is coordinated with a contrastive discourse relation; this is exemplified in (11).

(11) " I do n't know **why** , but I like Jimmy Carter . "

Three features are designed to capture this pattern: COORDWITHSLUICE indicates whether the sluice and candidate are connected by a coordination dependency, AFTERINITIALSLUICE marks the conjunctive condition where the candidate follows a sluice initial in its sentence, and IMMEDAFTERINITIALSLUICE marks a candidate that is the closest following candidate to an initial sluice.

[2] This feature might be seen as an analog to the apposition features used in nominal coreference resolution (Bengtson and Roth, 2008), but there it is used to link appositives, whereas here it is to exclude candidates.

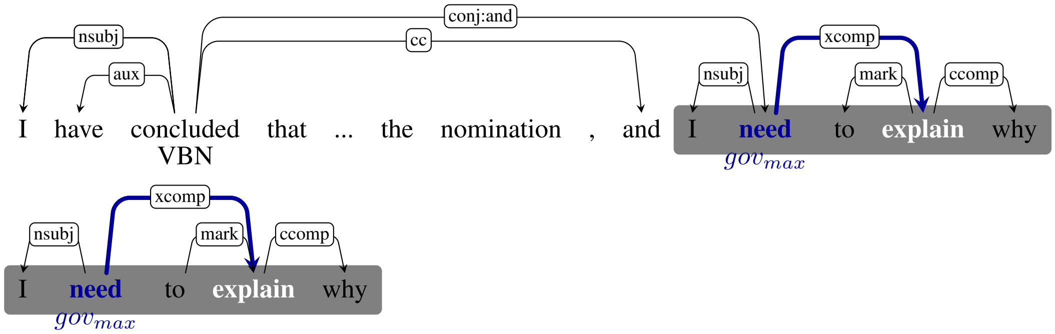

Figure 1: Sluice containment for S-4 and S-1 in (7). Starting at the governor of the sluice, *explain*, find gov$_{max}$ *need* and delete its transitive dependents. The candidate does not contain the sluice if the remaining graph contains verbal governors.

5 Content

In addition to the structural features above, we also compute several features relating the content of the sluice site and the antecedent. The intuition behind these relational features is the following: each sluice type (*why*, *who*, *how much*, etc.) represents a certain type of question, and each candidate represents a particular type of predication. For a given a sluice type, some predications might fit more naturally than others. More generally, it is a common view that an elliptical expression and its antecedent contain matching "parallel elements".[3]

Below we describe three approaches to this: one simply looks for lexical overlap – words that occur both in the sluice expression and in the candidate. The second involves a more general notion of how a predication fits with a sluice type. To capture this, we gather co-occurrence counts of main verb and sluice types. The third approach compares potential correlates in candidates with the type of sluice.

5.1 Overlap

One potential candidate for overlap information is the presence of a correlate in the antecedent. However, 75% of of sluices involve WH-phrases that typically involve no correlate (e.g., *how, when, why*). The pertinent exception to this are extent sluices (ones where the remnant is *how (much|many|JJ)*), which have been argued to heavily favor a correlate (Merchant, 2001), such as (12) below (though see (13) for a counterexample).

(12) The 49ers are $[_{corr}$ very good] .
It 's hard to know **how good** because the Cowboys were the only team in the league who could test them .

We thus compute the number of tokens of OVERLAP between the content terms in the WH-phrase sluice (non-WH, non prepositional) and the entire antecedent.

5.2 Wh-Predicate

Even for correlate-less sluices, the WH-phrase must semantically cohere with the main predicate of the antecedent. Thus, in (13), S-3 is a more likely antecedent than S-2 because *increase* is more likely to take an implicit extent than *predict*. Although we could have consulted a lexically rich resource (e.g, VerbNet, FrameNet), our hope was that this general approach could carry over to less argument-specific combinations such as *how* with *complete* and *raise* in (14).

(13) $[_{S-3}$ Deliveries would increase as a result of the acquisition] , $[_{S-2}$ he predicted] , but $[_{S-1}$ he would not say by how much]

(14) $[_{S-4}$ $[_{S-3}$ Once the city and team complete a contract] , the Firebirds will begin to raise \$ 9 million] , $[_{S-2}$ team president Yount said] , $[_{S-1}$ but he would not say how] .

Our assumption is that some main predicates are more likely than others for a given sluice type, and we wish to gather data that reveals these probabilities. This is somewhat similar to the approach of Hindle and Rooth (1993), who gather probabilities

[3] This term is from Dalrymple et al. (1991); a similar general view about parallelism in ellipsis arises in many different theories, such as Prüst et al. (1994) and Asher (1993).

that reflect the association of verbal and nominal heads with prepositions to disambiguiate prepositional phrase attachment.

One way to collect these would be to use our sluicing data, which consists of a total of 2185 annotated examples. However, the probabilities of interest are not about sluicing *per se*. Rather, they are about how well a given predication fits with a given type of question. Thus instead of using our comparatively small set of annotated sluicing examples, we used overt WH-constructions in Gigaword to observe cooccurrences between question types and main predicates. To find overt WH-constructions, we extracted all instances where a WH-phrase is: a) a dependent (to exclude cases like *Who?*) and b) not at the right edge of a VP (to exclude sluices like *know who*, per Anand and McCloskey (2015)). To further ensure that we were not overlapping with our dataset, we did this only for the non=NYT subsets of Gigaword (i.e., AFP, APW, CNA, and LTW). This procedure generated 687,000 WH-phrase instances, and 79,753 WH-phrase-governor bigram types. From these bigrams, we calculated WH-PREDICATE, the normalized pmi of WH-phrase type and governor lemma in Annotated Gigaword.

5.3 Correlate Overlap

Twenty-two percent of our data has correlates, and these correlates should be discriminative for particular sluice types. For example, temporal (*when*) sluices have timespan correlates (e.g., *tomorrow*, *later*), while entity (*who/what*) sluices have individuals as correlates (e.g., *someone*, *a book*). We extracted four potential integer-valued correlate features from each candidate: LocativeCorr is the number of primarily locative prepositions (those with a locative MLE in The Preposition Project (Litowski and Hargraves, 2005)). EntityCorr is the number of nominals in the candidate that are indefinite (bare nominals or ones with a determiner relation to *a, an* and weak quantifiers (*some*, *many*, *much*, *few*, *several*).TemporalCorr is the number of lexical patterns in the candidate for TIMEX3 annotations in Timebank 1.2 (Pustejovesky et al., 2016). WhichCorr is the pattern for entities plus *or*.

distance	DISTANCE, FOLLOWS
containment	ContainsSluice isDominatedBySluice
discourse structure	coordWithSluice, afterInitialSluice, immedAfterInitialSluice, candInParenthetical, candInRelClause
content	OVERLAP, WH-PREDICATE
correlate	LocativeCorr, EntityCorr, TemporalCorr, WhichCorr

Table 1: Summary of features used in experiments.

6 Algorithms

Mention-pair coreference models reduce coreference resolution to two steps: a local binary classification, and a global resolution of coreference chains. We may see antecedent selection as a similar two-stage process: classification on the probability a given candidate is an antecedent, and then selection of the most likely candidate for a given sluice. As Denis and Baldridge (2008) note, one limitation of this approach is that the overall rank of the candidates is never directly learned. They instead propose to learn the *rank* of a candidate c for antecedent a, modeled as the log-linear score of a candidate across a set of coreference models m, $(exp \sum_j w_j m_j(c, a))$, normalized by the sum of candidate scores. We apply the same approach to our problem, viewing each feature in Table 1 as a model, and estimating weights for the features by hill-climbing. We begin by defining constructed baselines which are implemented by manually assigning weights. We then consider the results of a maxent classifier over the features. Finally, we determine the weights directly by hill-climbing with random restarts.

6.1 Manual Baselines

Random simply selects candidates at random. Clst chooses the closest candidate that starts before the sluice. This is done by assigning a weight of -1 to DISTANCE and -10 to FOLLOWING (to exclude the following candidate), and 0 to all other features. ClstBef chooses the closest candidate that entirely precedes the sluice (i.e., starts before and does not contain the sluice site). To construct ClstBef, we change the weight of CONTAINSSLUICE to -10, which means that candidates containing the sluice will never be chosen.

6.2 A maxent model

We trained a maxent classifier on the features in Table 1 for the binary antecedent-not antecedent task. With 10-fold cross-validation on the test set, the maxent model achieved an average accuracy on the binary antecedent task of 87.1 and an F-score of 53.8 (P=63.9, R=46.5). We then constructed an antecedent selector that chose the candidate with the highest classifier score.

6.3 Hill-Climbing

We define a procedure to hill-climb over weights in order to maximize ConAccuracy over the entire training set (maximizing TokF yielded similar results, and is not reported here). Weights are initialized with random values in the interval [-10,10]. At iteration i, the current weight vector is compared to alternatives differing from it by the current step size on one weight, and the best new vector is selected. For the results reported here, we performed 13 random restarts and exponential step size $10{*}i^{.5}$ (values that maximized performance on the DS).

7 Results

We performed 10-fold cross-validation over TS on the hill-climbed and maxent models above, producing average ConAccuracy and TokF as shown in Table 2, which also gives results of the three baselines on the entire dataset. The hill-climbed approach with all features substantially outperformed the baselines, achieving a ConAccuracy of 72.4%.

We investigated the performance of our hill-climbing procedure with ablation of several feature subsets. We ablated features by group, as in Table 1. Table 2 shows the results for using four groups and only one group, as well as the top two three group and two group combinations.

	A:Tr	F:Tr	A:Tes	F:Tes
HC-DCSNR	73.8	72.4	72.4	71.5
HC-CSNR	41.8	51.8	40.3	51.6
HC-DCSR	72.9	71.6	72.1	71.0
HC-DSNR	53.5	59.1	52.7	58.3
HC-DCNR	65.8	67.1	64.6	65.9
HC-DCSN	73.3	72.1	72.7	71.8
HC-DCS	72.7	71.5	72.5	71.3
HC-DCN	65.6	67.8	64.3	66.8
HC-DC	63.3	65.4	63.0	65.4
HC-DS	51.2	57.2	50.9	57.1
HC-D	41.6	51.6	41.5	51.6
HC-C	30.6	45.1	28.9	45.3
HC-S	30.7	43.0	27.0	42.0
HC-N	30.5	38.6	30.7	38.2
HC-R	23.6	35.9	22.2	33.1
Maxent	65.3	70.2	64.2	68.0
Random	19.4	44.1	19.5	46.3
Clst	41.2	52.1	na	na
ClstBef	56.5	67.9	na	na

Table 2: Average (Con)A(ccuracy) and (Tok)F(-Score) for Tr(ain) and Tes(t) splits on 10-fold cross-validation of data. Feature groups: **D**istance, **C**ontainment, Discourse **S**tructure, co**N**tent, co**R**relate. (Red marks results not significantly different (via paired t-test) from HC-DCSNR.)

Features fall in three tiers. Distance features are the most predictive: all the top systems use them, and they alone perform reasonably well (like Clst). Containment and then Discourse Structure features are the next most helpful. The full system has a ConAccuracy of 72.4 on the TS, not reliably different from several systems without Content and/or Correlate features. At the same time, the scores for these feature types on their own show that they are predictive of the antecedent: The Correlate feature **R** has a score of 22.2, which is a rather modest, but statistically significant, improvement over Random. The Content feature **N** improves quite substantially, up to 30.7. This suggests that there is some redundancy with the other features, so that the contributions of Content and Correlate are not observed in combination with them. (HC-N and HC-R's lower than Random TokF is a result of precision: Random more often selects very small candidates inside the correct antecedent, leading to a higher precision.)

The Content and Correlate features concern relations between the type of sluice and the content of the antecedent; since other features do not capture this, it is puzzling that these provide no further improvement. To better understand why this is, we investigated the performance of our feature sets by sluice type. For the top performing systems, we found that antecedent selection for sluices over extents (e.g, *how much*, *how tall*) performed 11% better than average and those over reasons (*why*) and manner (*how*) performed 13% worse than average; no other WH-phrase types differed significantly from average. Importantly, this finding was consistent even for the systems without Content or Correlate features, which we extracted in large part to help highlight possible correlate material for extent sluices as well as entity (*who/what*) and temporal (*when*) sluices.

We also examined systems knocking out our best performing features, Distance, Containment, and Discourse Structure. When Distance features were omitted, we saw a bimodal distribution: reason and manner sluice antecedent selection was 31% better than expected (based on the full system differences discussed above), and the other sluices performed 22% worse. When Containment features were omitted, reason sluices performed 10% better than expected, while extent ones were 10% worse. Finally, when Discourse Structure features were removed, entity and temporal sluices had half the error rate we would expect. While it is hard to provide a clear takeaway from these differences, they do point to the relative difficulty in locating sluice antecedents based on WH-phrase type, and they also suggest that different sluice types present quite different challenges. This suggests that one promising line might be to learn different featural weights for each sluice type.

8 Conclusion

We have addressed the problem of sluicing antecedent selection by defining linguistically sophisticated features describing the structure and content of candidates. We described a hill-climbed model which achieves accuracy of 72.4%, a substantial improvement over a strong manually constructed baseline. We have shown that both syntactic and discourse relationships are important in antecedent selection. In future work, we hope to improve the performance of several of our features. Notable among these are the discourse structural proxies we found to make a contribution to the model. These features constitute a quite limited view of discourse structure, and we suspect that a better representation of discourse structure might well lead to further improvements. One potential path would be to leverage data where discourse relations are explicitly annotated, such as that in the Penn Discourse Treebank (Prasad et al., 2008). In addition, although our Content and Correlate features were not useful alongside the others, we hope that more refined versions of those could provide some assistance. We also noted that our performance was impacted by WH-types, and therefore it might be helpful to learn different featural weights per type.

In closing, we would like to return to the larger question of effectively handling ellipsis. The solution to antecedent selection that we have presented here provides a starting point for addressing the problem of *resolution*, in which the content of the sluice is filled in. However, even if the correct antecedent is selected, the missing content is not always an exact copy of the antecedent – often substantial modifications will be required – and an effective resolution system will have to negotiate such mismatches. As it turns out, many incorrect antecedents differ from the correct antecedent in ways highly reminiscent of these mismatches. Thus, some of the errors of our selection algorithm may be most naturally addressed by the resolution system, and it may be that the relative priority of the specific challenges we identified here will become clearer as we address the next step down in the overall pipeline.

Acknowledgments

We gratefully acknowledge the work of Jim McCloskey in helping to create the dataset investigated here, as well as the principal annotators on the project. We thank Jordan Boyd-Graber, Ellen Riloff, and three incisive reviewers for helpful comments. This research has been sponsored by NSF grant number 1451819.

References

Pranav Anand and Jim McCloskey. 2015. Annotating the implicit content of sluices. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 178.

Scott AnderBois. 2014. The semantics of sluicing: Beyond truth conditions. *Language*, 90(4):887–926.

Nicholas Asher. 1993. *Reference to Abstract Objects in English*. Dordrecht.

Henry Beecher. 2008. Pramatic inference in the interpretation of sluiced Prepositional Phrases. In *San Diego Linguistic Papers*, volume 3, pages 2–10. Department of Linguistics, UCSD, La Jolla, California.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Johan Bos and Jennifer Spenader. 2011. An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463–494.

Sara Cantor. 2013. Ungrammatical double-island sluicing as a diagnostic of left-branch positioning.

S. Chung, W. Ladusaw, and J. McCloskey. 1995. Sluicing and logical form. *Natural Language Semantics*, 3:1–44.

Mary Dalrymple, Stuart Shieber, and Fernando Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy*, 14(4), August.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 660–669.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2005. Automatic bare sluice disambiguation in dialogue. In *Proceedings of the IWCS-6 (Sixth International Workshop on Computational Semantics)*, pages 115–127, Tilburg, the Netherlands, January. Available at: `http://www.dcs.kcl.ac.uk/staff/lappin/recent_papers_index.html`.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Raquel Fernández, Jonathan Ginzburg, Howard Gregory, and Shalom Lappin. 2008. Shards: Fragment resolution in dialogue. In *Computing Meaning*, pages 125–144. Springer.

Jonathan Ginzburg and Ivan Sag. 2000. *Interrogative Investigations: The Form, Meaning and Use of English Interrogatives*. CSLI Publications, Stanford, Calif.

Daniel Hardt and Maribel Romero. 2004. Ellipsis and the structure of discourse. *Journal of Semantics*, 24(5):375–414.

Daniel Hardt. 1997. An empirical approach to vp ellipsis. *Computational Lingusitics*, 23(4):525–541.

Daniel Hardt. 1999. Dynamic interpretation of verb phrase ellipsis. *Linguistics & Philosophy*, 22(2):187–221.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational linguistics*, 19(1):103–120.

Varada Kolhatkar and Graeme Hirst. 2012. Resolving "this-issue" anaphora. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1255–65.

Varada Kolhatkar, Heike Zinmeister, and Graeme Hirst. 2013. Interpreting anaphoric shell nouns using antecedents of ctaphoric shell nouns as training data. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Ken Litowski and Orin Hargraves. 2005. The preposition project. In *ACL-SIGSEM Workshop on "The Linguistic Dimension of Prepositions and Their Use in Computational Linguistic Formalisms and Applications*, pages 171–179.

Anne Lobeck. 1995. *Ellipsis: Functional heads, licensing and identification*. Oxford University Press.

Jason Merchant. 2001. *The syntax of silence: Sluicing, islands, and identity in ellipsis*. Oxford.

Leif Nielsen. 2005. *A Corpus-Based Study of Verb Phrase Ellipsis Identification and Resolution*. Ph.D. thesis, King's College London.

Johanna Nykiel. 2010. Whatever happened to Old English sluicing. In Robert A. Cloutier, Anne Marie Hamilton-Brehm, and Jr. William A. Kretzschmar, editors, *Studies in the History of the English Language V: Variation and Change in English Grammar and Lexicon: Contemporary Approaches*, pages 37–59. Walter de Gruyter.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Hub Prüst, Remko Scha, and Martin van den Berg. 1994. A discourse perspective on verb phrase anaphora. *Linguistics and Philosophy*, 17(3):261–327.

James Pustejovesky, Marc Verhagen, Roser Sauri, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. 2016. Timebank 1.2. LDC2006T08, April.

Ivan A. Sag. 1976. *Deletion and Logical Form*. Ph.D. thesis, Massachusetts Institute of Technology. (Published 1980 by Garland Publishing, New York).

W. M. Soon, H.T. Ng, and D. C. Y Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–44.