Classifying Sluice Occurrences in Dialogue

Austin Baird, Anissa Hamza, Daniel Hardt
UCSC, LiLPa/Unistra, Copenhagen Business School/UCSC
arbaird@ucsc.edu, anissa.hamza@etu.unistra.fr, dhardt@ucsc.edu

Abstract

Ellipsis is an important challenge for natural language processing systems, and addressing that challenge requires large collections of relevant data. The dataset described by Anand and McCloskey (2015), consisting of 4100 occurrences, is an important step towards addressing this issue. However, many NLP technologies require much larger collections of data. Furthermore, previous collections of ellipsis are primarily restricted to news data, although sluicing presents a particularly important challenge for dialogue systems. In this paper we classify sluices as Direct, Reprise, Clarification. We perform manual annotation with acceptable inter-coder agreement. We build classifier models with Decision Trees and Naive Bayes, with accuracy of 67%. We deploy a classifier to automatically classify sluice occurrences in OpenSubtitles, resulting in a corpus with 1.7 million occurrences. This will support empirical research into sluicing in dialogue, and it will also make it possible to build NLP systems using very large datasets. This is a noisy dataset; based on a small manually annotated sample, we found that only 80% of instances are in fact sluices, and the accuracy of sluice classification is lower. Despite this, the corpus can be of great use in research on sluicing and development of systems, and we are making the corpus freely available on request. Furthermore, we are in the process of improving the accuracy of sluice identification and annotation for the purpose of created a subsequent version of this corpus.

Keywords: sluicing, ellipsis, dialogue

1. Introduction

Ellipsis is a major challenge for NLP systems, as well as an important topic in theoretical linguistics. The most extensive empirical work to date on ellipsis is described in Anand and Hardt (2016) and Anand and McCloskey (2015). This work involves a corpus of some 4100 sluice occurrences, extracted from the NYTimes Gigaword Corpus. These occurrences have been manually annotated in a detailed fashion.

Sluices are elliptical questions, where all but the interrogative phrase of a question is omitted, leaving a wh-word remnant, as in the following example, with the sluice wh-word in bold (Anand and Hardt, 2016):

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

In this paper, we construct a very large corpus of sluice occurrences in dialog. We build on previous work (Anand and McCloskey, 2015; Fernández et al., 2004; Fernández et al., 2007) in developing methods to automatically identify and classify sluice occurrences. We apply these methods to the English portion of OpenSubtitles, resulting in a corpus of over 1.7 million sluice occurrences. This is orders of magnitude larger that any previous collections of ellipsis occurrences and it has been automatically annotated with linguistically relevant features.

2. Related Work

(Fernández et al., 2004; Fernández et al., 2007) describe an approach to the classification of sluice occurrences in the British National Corpus (BNC). Fernandez et al. focus on what they call bare sluices: utterances in dialog consisting of only a wh-word (they also consider the form *which N*). They extract 5343 bare sluices from the dialogue transcripts of the BNC. Fernández et al. (2004) classify dialogue sluices as follows.

Feature	Description
sluice	type of sluice
mood	declarative or non-declarative
polarity	positive or negative
frag	fragment or not
quant	presence of a quantified expression
deictic	presence of a deictic pronoun
proper_n	presence of a proper name
pro	presence of a pronoun
def_desc	presence of a definite description
Wh	presence of a wh-word
overt	presence of other potential antecedent expression

Table 1: Features

Direct: the sluice queries for additional information that was explicitly or implicitly quantified away in the previous utterance.

Reprise: The utterer of the sluice cannot understand some aspect of the previous utterance which the previous speaker assumed as presupposed.

Clarification: the sluice used to ask for clarification about the previous utterance as a whole.

Wh-anaphor: the antecedent is a *wh*-phrase.

(They also use a category **Unclear**, which we will ignore.) Fernandez et al. build models to classify sluice occurrences, using the above five-way classification scheme. They define the features as shown in Table 1: the first is the type of sluice; the other features all apply to the antecedent utterance.

A total of 351 data points were used to train the classifiers. Table 2 gives the distribution of these data points by the

Sluice	Direct n (%)	Reprise n (%)	Clarification n (%)	Wh-anaphor n (%)
What	7 (9.60)	17 (23.3)	17 (23.3)	1 (1.3)
Why	55 (68.7)	24 (30.0)	0 (0)	1 (1.2)
Who	10 (13.0)	65 (84.4)	0 (0)	2 (2.6)
Where	31 (34.4)	56 (62.2)	0 (0)	3 (3.3)
When	50 (63.3)	27 (34.1)	0 (0)	2 (2.5)
Which	1 (8.3)	11 (91.6)	0 (0)	0 (0)
whichN	19 (21.1)	71 (78.8)	0 (0)	0 (0)
How	23 (79.3)	3 (10.3)	3 (10.3)	0 (0)
Total	106(30.2)	203(57.8)	24 (6.8)	18 (5.1)

Table 2: Sluice Cats and Wh Types

Reading	Recall	Precision	F1
Direct	71.70	79.20	75.20
Reprise	85.70	83.70	84.70
Clarification	100.00	68.60	81.40
Wh anaphor	66.70	100.00	80.00
weighted score	81.47	82.14	81.80

Table 3: BNC Sluice Classification

wh-word and classification (Fernández et al. (2007), table 3).

Four machine learning classifiers were run on this dataset annotated with the 11 features, with weighted f-scores ranging from 73.24 - 81.62. Table 3 shows the results obtained by the most accurate learner (Fernández et al., 2007) (Appendix A)

3. The Data

3.1. Opensubtitles

The English portion of Opensubtitles (http://www.opensubtitles.org/)[1] contains 2,125,277,188 words and 327,968,003 lines. Building on methods described in (Anand and McCloskey, 2015), we locate both root sluices and embedded sluices. As explained in (Anand and McCloskey, 2016) a root sluice is unembedded (2), while non-root sluices are "sub-parts of larger structures", as in (3).

(2) A: We should go home. B: Why/when/what for/how?

(3) The university has to change, but it's not clear in what ways.

In order to locate sluices, the entire corpus was first POS tagged using the Stanford POS tagger, described in (Toutanova et al., 2003). We define two regular expressions to search for sluices in the corpus. The first identifies embedded sluices with a pattern including an embedding verb

[1]Pierre Lison and Jörg Tiedemann, 2016, OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)

WH word	root count	embedded count
What	1,097,382	14,421
Why	352,047	29,805
How	122,256	6,453
Who	98,330	2,335
Where	70,312	2,677
When	29,171	1,473
Which	18,491	308
Whom	8,874	60

Table 4: Root vs. embedded sluice in Opensubtitles

followed by a wh-word. The wh-word is optionally followed by an adjective, adverb, preposition, or noun. Following this is an optional punctuation, followed by end-of-string. This pattern defines the following list of embedding verbs: *know, knew, ask, say, understand, wonder, remember, tell, explain, imagine, care, forget, worry*. The second pattern applies to lines that did not match the first pattern. It is the same as the one described above, except that it does not contain an embedding verb and the sentence must end with a '?'.
Of the sluices found in the corpus, a total of 57,532 were embedded sluices and a total of 1,796,863 were root sluices. Table 4 gives the breakdown of root and embedded sluices by wh word.

3.2. Annotating Sluice Types

We construct two samples of sluice occurrences for the purpose of manual annotation: the first includes the first 100 root sluices. Since the distribution of wh-words is quite unbalanced, we construct a more balanced sub-corpus, which includes 1000 randomly selected examples of what, and 500 of each remaining wh-word – why, how, who, where, when, which, whom – making up a total of 4500 examples. We used four categories, following Fernández et al. (2007), with the following revised definitions:

Direct questions an indefinite part of the antecedent that is implicitly or explicitly expressed, and is not necessarily known by the speaker.
A: He didn't come.
B: **Why?**
A: Break up.

Clarification questions the entire antecedent, typically expressing surprise or confusion.
A: Captain ! It 's the Tomb of Heroes !
B: **What?**
How can it be?

It also includes illocutionary uses of *wh*-words as in:
A: Congratulations on your promotion !
B: Should I thank you ?
A: **Why ?**

Sluices lacking a linguistic antecedent are also classified under Clarification.
A: It 's Colonel Gelovani.

Sluice Type	Agreement
Clarification	87.9%
Direct	83.8%
Reprise	61.5%

Table 5: Intercoder Agreement for Opensubtitles Sluices

B: Yes .
A: **What ?**

Reprise addresses a definite and explicit part of the antecedent. The questioned element is definitely known to the speaker.
A: They made her mad.
B: **Who ?**
A: The devils

None occurrences are not in fact sluices. These are often due to incorrect POS labels, or frozen questions which typically occur in spontaneous and oral discourse.
A : She teases him a lot .
B : That 's natural in a girl
A : Yes , I suppose so . What about Claudius ?

3.3. Interannotator Agreement

The three authors of this paper individually annotated the same sample of 100 sluices, resulting in 84% average agreement. The kappa score is 80%.
Out of the sample, 51% of the sluices were unanimously classified as Clarification, 8% as Reprise, 26% as direct, and 3% as None. The agreement rates for sluices are given in table 5. (The agreement rate for None was 100%)
Although Reprise sluices were the least frequent in the sample (besides None), they had the highest amount of disagreement. In all disagreements involving a Reprise sluice, the alternative classification by the disagreeing annotator was Direct. For all disagreements involving a Clarification sluice, the disagreeing annotator always annotated as Direct as well. These disagreements overwhelmingly occurred in sluices containing a single 'what?', where the preceding and succeeding context was needed in order to determine the type. All instances in which all three annotators disagreed on the sluice type are not included for the percentage calculations. Due to the relatively high overall agreement among the authors, a single author annotated all the samples used in training the classifier for this paper.

4. Predictive Model

4.1. Training Data

Two sets of training data were used in building classifiers, as shown in Table 6. Set2 roughly matches the distribution of classes in OpenSubtitles, while Set1 is more balanced.
The decision tree classifiers are built using scikit-learn (Pedregosa et al., 2011), and the Naive Bayes classifiers are using nltk (Bird et al., 2009).
The features used to train the were identical to nine of the features used by the authors of Fernández et al. (2007). All of the features described in section 2 are used except for frag and overt. All of the features, except for 'type', take on boolean values. The value for 'type' is the wh word contained in the sluice. Unlike the features in Fernández et al. (2007), there is no distinction between WhichN and Which for the classifier used in this paper.
The separate datasets were used to train both a NaiveBayes classifier and a Decision Tree classifier. Both classifiers have accuracies scored using 10-fold cross validation. In what follows, we focus on the Decision Tree classifier results on the balanced dataset, Set1, as these were the best results.

	Set 1					Set 2				
	C	D	R	N	Total	C	D	R	N	Total
What	674	126	54	107	961	711	126	54	109	1000
Why	126	224	109	4	463	68	203	77	2	350
How	0	380	28	15	423	0	105	13	2	120
Who	0	40	191	45	276	0	28	63	9	100
Where	0	20	86	66	172	0	47	14	9	70
When	0	0	56	30	86	0	20	8	2	30
Which	0	10	33	35	78	0	1	15	4	20
Whom	0	0	43	30	73	0	9	0	1	10
Total	800	800	600	332	2532	779	539	244	136	1700

Table 6: Two Training Datasets

class	Precision	Recall	f1-score
clar	0.71	0.91	0.80
dir	0.80	0.81	0.81
rep	0.86	0.74	0.79
none	0.85	0.53	0.65
avg / total	0.80	0.75	0.78

Table 7: Decision Tree Set1

4.2. Classifier Results

Table 7 shows the results using the decision tree classifier with Set1. The majority baseline results are shown in Table 8.
This classifier beats the majority baseline overall and performs relatively well in most areas. However, it has a very low recall when identifying None type sluices. We suspect that this is because other features are relevant to identifying this class.

5. Classifying All OpenSubtitles Sluices

The decision tree classifier was used to classify all of the sluices detected in OpenSubtitles. The number of classifications assigned to the sluices are shown in Table 9. The

class	Precision	Recall	f1-score
clar	0.31	1.00	0.47
dir	0.00	0.00	0.00
rep	0.00	0.00	0.00
none	0.00	0.00	0.00
avg / total	0.10	0.31	0.15

Table 8: Baseline Set1

Class	Amount	Percentage
Clarification	1,110,210	61.8%
Direct	379,420	21.1%
Reprise	226,568	12.6%
None	80,665	4.5%

Table 9: Resulting Dataset

	Clar	Dir	Rep	None
What	1,059,912	13,578	2,947	19,759
Why	50,298	232,768	69,006	98
How	0	120,498	1,411	422
Who	0	2,514	87,911	7,938
Where	0	7,188	29,120	34,033
When	0	0	21,674	7,539
Which	0	2,874	8,539	7,088
Whom	0	0	5,960	3,788
Total	1,110,210	379,420	226,568	80,665

Table 10: Class by wh-word in OpenSubtitles

type of sluice as a percentage of all sluices detected are also shown in this table. Note that the total number of sluices includes those that the classifier classified as None.

Table 9 gives the resulting dataset, broken down by class; this is further broken down by wh-word in Table 10.

A random sample of 103 examples classified by the model were selected and hand annotated to compute the classifier's accuracy on this sample. Table 11 shows two sets of percentages about the classifier's predictions. First, of all the sluices in categories Direct, Clarification, and Reprise, the percentage of which are actually sluices (not annotated as being of the None class). Second, of the sluices categorized as Direct, Clarification, or Reprise, what percent are correct.

Table 11 shows that overall, 80% of the examples that the classifier predicted to be a sluice were actually sluices, and 67% were categorized correctly by the classifier.

6. Conclusion

Ellipsis is an important challenge for natural language processing systems, and addressing that challenge requires large collections of relevant data. The dataset described by Anand and McCloskey (2015), consisting of 4100 occurrences, is an important step towards addressing this issue. However, many NLP technologies require much larger collections of data. Furthermore, previous collections of ellipsis are primarily restricted to news data, although sluicing presents a particularly important challenge for dialogue systems.

Predicted Class	True Sluices	Correctly Categorized
clar	0.81	0.76
dir	0.69	0.61
rep	0.57	0.57
Total	0.80	0.67

Table 11: Classifier Accuracy Results

In this paper we present an ellipsis corpus with 1.7 million occurrences. This will support empirical research into sluicing in dialogue, and it will also make it possible to build NLP systems using very large datasets. This is a noisy dataset; based on a small manually annotated sample, we found that only 80% of instances are in fact sluices, and the accuracy of sluice classification is lower. Despite this, the corpus can be of great use in research on sluicing and development of systems, and we are making the corpus freely available on request. Furthermore, we are in the process of improving the accuracy of sluice identification and annotation for the purpose of created a subsequent version of this corpus.

7. Acknowledgments

Thanks to three reviewers for helpful comments. We gratefully acknowledge Pranav Anand and Jim McCloskey for their work on sluicing annotation and analysis on the project *The Implicit Content of Sluicing*, which provides the inspiration and background for the current work. This research has been sponsored by NSF grant 1451819.

8. Bibliographical References

Anand, P. and Hardt, D. (2016). Antecedent selection for sluicing: Structure and content. In *EMNLP*, pages 1234–1243.

Anand, P. and McCloskey, J. (2015). Annotating the implicit content of sluices.

Anand, P. and McCloskey, J. (2016). Annotation guide.

Fernández, R., Ginzburg, J., and Lappin, S. (2004). Classifying ellipsis in dialogue: A machine learning approach. In *Proceedings of the 20th international conference on computational linguistics*, page 240. Association for Computational Linguistics.

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

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL 2003*, pages 252–259.

9. Language Resource References

Bird, Steven and Loper, Edward and Klein, Ewan. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*.