Annotating Anaphoric Shell Nouns with their Antecedents

Varada Kolhatkar
Department of Computer Science
University of Toronto
varada@cs.toronto.edu

Heike Zinsmeister
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
zinsmeis@ims.stuttgart-uni.de

Graeme Hirst
Department of Computer Science
University of Toronto
gh@cs.toronto.edu

Abstract

Anaphoric shell nouns such as *this issue* and *this fact* conceptually encapsulate complex pieces of information (Schmid, 2000). We examine the feasibility of annotating such anaphoric nouns using crowdsourcing. In particular, we present our methodology for reliably annotating antecedents of such anaphoric nouns and the challenges we faced in doing so. We also evaluated the quality of crowd annotation using experts. The results suggest that most of the crowd annotations were good enough to use as training data for resolving such anaphoric nouns.

1 Introduction

Anaphoric shell nouns (ASNs) such as *this fact*, *this possibility*, and *this issue* are common in all kinds of text. They are called *shell nouns* because they provide nominal conceptual shells for complex chunks of information representing abstract concepts such as *fact*, *proposition*, and *event* (Schmid, 2000). An example is shown in (1).

(1) Despite decades of education and widespread course offerings, **the survival rate for out-of-hospital cardiac arrest remains a dismal 6 percent or less worldwide**.

This fact prompted the American Heart Association last November to simplify the steps of CPR to make it easier for lay people to remember and to encourage even those who have not been formally trained to try it when needed.

Here, the ASN *this fact* encapsulates the clause marked in bold from the preceding paragraph.

ASNs play an important role in organizing a discourse. First, they are used metadiscursively to talk about the current discourse. In (1), the author *characterizes* the information presented in the context by referring to it as a *fact* — a thing that is indisputably the case. Second, they are used as cohesive devices in a discourse. In (1), for example, *this fact* on the one hand refers to the proposition marked in bold, and on the other, faces forward and serves as the starting point of the following paragraph. Finally, as Schmid (2000) points out, like conjunctions *so* and *however*, ASNs may function as topic boundary markers and topic change markers.

Despite their importance, ASNs have not received much attention in Computational Linguistics. Although there has been some effort to annotate certain anaphors with similar properties, i.e., demonstratives and the pronoun *it* (Byron, 2003; Artstein and Poesio, 2006), in contrast to ordinary nominal anaphora, there are not many annotated corpora available that could be used to study ASNs. Indeed, many questions of annotation of ASNs must still be answered. For example, the extent to which native speakers themselves agree on the resolution of such anaphors, i.e., on the precise antecedents, remains unclear.

An essential first step in this field of research is therefore to clearly establish the extent of inter-annotator agreement on antecedents of ASNs as a measure of feasibility of the task. In this paper, we describe our methodology for annotating ASNs using crowdsourcing, a cheap and fast way of obtaining annotation. We also describe how we evaluated the feasibility of the task and the quality of the annotation, and the challenges we faced in doing so, both with regard to the task itself and the crowdsourcing platform we use. The results suggest that most of the crowd-annotations were good enough to use as training data for ASN resolution.

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 112–121,
Sofia, Bulgaria, August 8-9, 2013.

2 Related work

There exist only few annotated corpora of anaphora with non-nominal antecedents (Dipper and Zinsmeister, 2011). The largest one of these, the ARRAU corpus (Poesio and Artstein, 2008), contains 455 anaphors pointing to non-nominal antecedents, but only a few instances are ASNs. Kolhatkar and Hirst (2012) annotated antecedents of the same type as we do, but restricted their efforts to the ASN *this issue*.[1] In addition, there are corpora annotated with event anaphora in which verbal instances are identified as proxies for non-nominal antecedents (Pradhan et al., 2007; Chen et al., 2011; Lee et al., 2012).

For the task of identifying non-nominal antecedents as free spans of text, there is no standard way of reporting inter-annotator agreement. Some studies report only observed percentage agreement with results in the range of about 0.40–0.55 (Vieira et al., 2002; Dipper and Zinsmeister, 2011). The studies differed with respect to number of annotators, types of anaphors, and language of the corpora. Artstein and Poesio (2006) discuss Krippendorff's alpha for chance-corrected agreement. They considered antecedent strings as bags of words and computed the degree of difference between them by different distance measures (e.g. Jaccard, Dice). The bag-of-words approach is rather optimistic in the sense that even two non-overlapping strings are very likely to share at least a few words. Kolhatkar and Hirst (2012) followed a different approach by using Krippendorff's unitizing alpha ($_u\alpha$) which considers the longest common subsequence of different antecedent options (Krippendorff, 2013). They reported high chance-corrected $_u\alpha$ of 0.86 for two annotators but in a very restricted domain.

There has been some prior effort to annotate anaphora and coreference using *Games with a Purpose* as a method of crowdsourcing (Chamberlain et al., 2009; Hladká et al., 2009). Another, less time-consuming approach of crowdsourcing is using platforms such as Amazon Mechanical Turk[2]. It has been shown that crowdsourced data can successfully be used as training data for NLP tasks (Hsueh et al., 2009).

[1] Another data set reported in the literature could have been relevant for us: Botley's (2006) corpus contained about 462 ASN instances signaled by shell nouns; but this data is no longer available (S. Botley, p.c.).

[2] `https://mturk.com/mturk/`

Class	Description	Examples
factual	states of affairs	***fact***, ***reason***
linguistic	linguistic acts	***question***, *report*
mental	thoughts and ideas	***issue***, ***decision***
modal	subjective judgements	***possibility***, *truth*
eventive	events	*act*, *reaction*
circumstantial	situations	*situation*, *way*

Table 1: Schmids categorization of shell nouns. The nouns in boldface are used in this research.

3 The Anaphoric Shell Noun Corpus

Our goal is to obtain annotated data for ASN antecedents that could be used to train a supervised machine learning system to resolve ASNs. For that, we created the **A**naphoric **S**hell **N**oun (ASN) corpus.

Schmid (2000) provides a list of 670 English nouns which are frequently used as shell nouns. He divides them into six broad semantic classes: *factual*, *mental*, *linguistic*, *modal*, *circumstantial*, and *eventive*. Table 1 shows this classification, along with example shell nouns for each category.

To begin with, we considered articles containing occurrences of these 670 shell nouns from the New York Times (NYT) corpus (about 711,046 occurrences).[3] To create a corpus of a manageable size for annotation, we considered first 10 highly frequent shell nouns distributed across each of Schmid's shell noun categories from Table 1 and extracted ASN instances by searching for the pattern {*this* shell_noun} in these articles.[4]

To examine the feasibility of the annotation, we systematically annotated sample data ourselves, which contained about 15 examples of each of these 10 highly frequent shell nouns. The annotation process revealed that not all ASN instances are easy to resolve. The instances with shell nouns from the circumstantial and eventive categories, in particular, had very long and unclear antecedents. So we excluded these categories in this research and work with six shell nouns from the other four categories: *fact*, *reason*, *issue*, *decision*, *question*, and *possibility*. To create the ASN corpus, we extracted about 500 instances for each of these six shell nouns. After removing duplicates and instances with a non-abstract sense (e.g., *this is-*

[3] `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T19`

[4] Schmid (2000) provides patterns for anaphoric shell nouns, and *this*-NP is the most prominent pattern among them.

sue with a publication-related sense), we were left with 2,822 ASN instances.

4 ASN Annotation Challenges

ASN antecedent annotation is a complex task, as it involves deeply understanding the discourse and interpreting it. Here we point out two main challenges associated with the task.

What to annotate? The question of 'what to annotate' as mentioned by Fort et al. (2012) is not straightforward for ASN antecedents, as the notion of *markables* is complex compared to ordinary nominal anaphora: the units on which the annotation work should focus are heterogeneous.[5] Moreover, due to this heterogeneous nature of annotation units, there is a huge number of markables (e.g., all syntactic constituents given by a syntactic parse tree). So there are many options to choose from, while only a few units are actually to be annotated. Moreover, there is no one-to-one correspondence between the syntactic type of an antecedent and the semantic type of its referent (Webber, 1991). For instance, a semantic type such as *fact* can be expressed with different syntactic shapes such as a clause, a verb phrase, or a complex sentence. Conversely, a syntactic shape, such as a clause, can function as several semantic types, including *fact*, *proposition*, and *event*.

Lack of the notion of the *right* answer It is not obvious how to define clear and detailed annotation guidelines to create a gold-standard corpus for ASN antecedent annotation due to our limited understanding of the nature and interpretation of such antecedents. The notion of the *right* answer is not well-defined for ASN antecedents. Indeed most people will be hard-pressed to say whether or not to include the clause *Despite decades of education and widespread course offerings* in the antecedent of *this fact* in example (1). The main challenge is to identify the conditions when two different candidates for annotation should be considered as representing essentially the same concept, which raises deep philosophical issues that we do not propose to solve in this paper. For our purposes, we believe, this challenge could only be possibly tackled by the requirements of downstream applications of ASN resolution.

[5]Occasionally, ASN antecedents are non-contiguous spans of text, but in this work, we ignore them for simplicity.

5 Annotation Methodology

Considering the difficulties of ASN annotation discussed above, there were two main challenges involved in the annotation process: first, to find annotators who can annotate data reliably with minimal guidelines, and second, to design simple annotation tasks that will elicit data useful for our purposes. Now we discuss how we dealt with these challenges.

Crowdsourcing We wanted to examine to what extent non-expert native speakers of English with minimal annotation guidelines would agree on ASN antecedents. We explored the possibility of using *crowdsourcing*, which is an effective way to obtain annotations for natural language research (Snow et al., 2008). In particular, we explored the use of CrowdFlower[6], a crowdsourcing platform that in turn uses various worker channels such as Amazon Mechanical Turk. CrowdFlower offers a number of features.

First, it offers a number of integrated *quality-control* mechanisms. For instance, it throws gold questions randomly at the annotators, and annotators who do not answer them correctly are not allowed to continue. To further minimize spammers, it also offers a training phase before the actual annotation. In this phase, every annotator is presented with a few gold questions. Only those annotators who get the gold questions right get admittance to do the actual annotation.

Second, CrowdFlower chooses a unique answer for each annotation unit based on the majority vote of the trusted annotators. For each annotator, it assigns a trust level based on how she performs on the gold examples. The unique answer is computed by adding together the trust scores of annotators, and then picking the answer with the highest sum of trusts (CrowdFlower team, p.c.). It also assigns a *confidence* score (denoted as c henceforth) for each answer, which is a normalized score of the summation of the trusts. For example, suppose annotators A, B, and C with trust levels 0.75, 0.75, and 1.0 give answers *no*, *yes*, *yes* respectively for a particular instance. Then the answer *yes* will score 1.75 and answer *no* will score 0.75 and *yes* will be chosen as the crowd's answer with $c = 0.7$ (i.e., $1.75/(1.75 + 0.75)$). We use these confidence scores in our analysis of inter-annotator agreement below.

[6]`http://crowdflower.com/`

Finally, CrowdFlower also provides detailed annotation results including demographic information and trustworthiness of each annotator.

Design of the annotation tasks With the help of well-designed gold examples, CrowdFlower can get rid of spammers and ensures that only reliable annotators perform the annotation task. But the annotation task must be well-designed in the first place to get a good quality annotation. Following the claim in the literature that with crowdsourcing platforms simple tasks do best (Madnani et al., 2010; Wang et al., 2012), we split our annotation task into two relatively simple sequential annotation tasks. First, identifying the broad region of the antecedent, i.e., not the precise antecedent but the region where the antecedent lies, and second, identifying the precise antecedent, given the broad region of the antecedent. Now we will discuss each of our annotation tasks in detail.

5.1 CrowdFlower experiment 1

The first annotation task was about identifying the broad region of ASN antecedents without actually pinpointing the precise antecedents. We defined the broad region as the sentence containing the ASN antecedent, as the shell nouns we have chosen tend to have antecedents that lie within a single sentence. We designed a CrowdFlower experiment where we presented to the annotators ASNs from the ASN corpus with three preceding paragraphs as context. Sentences in the vicinity of ASNs were each labelled: four sentences preceding the anaphor, the sentence containing the anaphor, and two sentences following the anaphor. This choice was based on our pilot annotation: the antecedents very rarely occur more than four sentences away from the anaphor. The annotation task was to pinpoint the sentence in the presented text that contained the antecedent for the ASN and selecting the appropriate sentence label as the correct answer. If no labelled sentence in the presented text contained the antecedent, we suggested to the annotators to select *None*. If the antecedent spanned more than one sentence, then we suggested to them to select *Combination*. We also provided a link to the complete article from which the text was drawn in case the annotators wanted to have a look at it.

Settings We asked for 8 judgements per instance and paid 8 cents per annotation unit. Our job contained in total 2,822 annotation units with 168 gold units. As we were interested in the verdict of native speakers of English, we limited the allowed demographic region to English-speaking countries.

5.2 CrowdFlower experiment 2

This annotation task was about pinpointing the exact antecedent text of the ASN instances. We designed a CrowdFlower experiment, where we presented to the annotators ASN instances from the ASN corpus with highlighted ASNs and the sentences containing the antecedents, the output of experiment 1. One way to pinpoint the exact antecedent string is to ask the annotators to mark free spans of text within the antecedent sentence, similar to Byron (2003) and Artstein and Poesio (2006). However, CrowdFlower quality-control mechanisms require multiple-choice annotation labels. So we decided to display a set of labelled candidates to the annotators and ask them to choose the answer that best represents the ASN antecedent. A practical requirement of this approach is that the number of options to be displayed be only a handful in order to make it a feasible task for online annotation. But as we noted in Section 4, the number of markables for ASN antecedents is large. If, for example, we define markables as all syntactic constituents given by the Stanford parser[7], there are on average 49.5 such candidates per sentence in the ASN corpus. It is not practical to display all these candidates and to ask CrowdFlower annotators to choose one answer from this many options. Also, some potential candidates are clearly not appropriate candidates for a particular shell noun. For instance, the NP constituent *the survival rate* in example (1) is not an appropriate candidate for the shell noun *fact* as generally facts are propositions. So the question is whether it is possible to restrict this set of candidates by discarding unlikely ones.

To deal with this question, we used supervised machine learning methods trained on easy, non-anaphoric unlabelled examples of shell nouns (e.g., *the fact that X*). In this paper, we will focus on the annotation and will treat these methods as a black box. In brief, the methods reduce the large search space of ASN antecedent candidates to a size that is manageable for crowdsourcing annotation, without eliminating the most likely candi-

[7] `http://nlp.stanford.edu/software/lex-parser.shtml`

dates. We displayed the 10 most-likely candidates given by these methods. In addition, we made sure not to display two candidates with only a negligible difference. For example, given two candidates, *X* and *that X*, which differ only with respect to the introductory *that*, we chose to display only the longer candidate *that X*.

In a controlled annotation, with detailed guidelines, such difficulties of selecting between minor variations could be avoided. However, such detailed annotation guidelines still have to be developed.

Settings As in experiment 1, we asked for 8 judgements per instance and paid 6 cents per annotation unit. But for this experiment we considered only 2,323 annotation units with 151 gold units, only high-confidence units ($c \geq 0.5$) from experiment 1. This task turned out to be a suitable task for crowdsourcing as it offered a limited number of options to choose from, instead of asking the annotators to mark arbitrary spans of text.

6 Agreement

Our annotation tasks pose difficulties in measuring inter-annotator agreement both in terms of the task itself and the platform used for annotation. In this section, we describe our attempt to compute agreement for each of our annotation tasks and the challenges we faced in doing so.

6.1 CrowdFlower experiment 1

Recall that in this experiment, annotators identify the sentence containing the antecedent and select the appropriate sentence label as their answer. We know from our pilot annotation that the distribution of such labels is skewed: most of the ASN antecedents lie in the sentence preceding the anaphor sentence. We observed the same trend in the results of this experiment. In the ASN corpus, the crowd chose the preceding sentence 64% of the time, the same sentence 13% of the time, and long-distance sentences 23% of the time.[8] Considering the skewed distribution of labels, if we use traditional agreement coefficients, such as Cohen's κ (1960) or Krippendorff's α (2013), expected agreement is very high, which in turn results in a low reliability coefficient (in our case $\alpha = 0.61$) that does not necessarily reflect the true reliability of the annotation (Artstein and Poesio, 2008).

[8]This confirms Passonneau's (1989) observation that non-nominal antecedents tend to be close to the anaphors.

	F	*R*	*I*	*D*	*Q*	*P*	*all*
$c < .5$	8	8	36	21	13	7	16
$.5 \leq c < .6$	6	6	13	8	7	5	8
$.6 \leq c < .8$	24	25	31	31	22	27	27
$.8 \leq c < 1.$	22	23	11	14	19	25	18
$c = 1.$	40	38	9	26	39	36	31
Average c	.83	.82	.61	.72	.80	.83	.76

Table 2: CrowdFlower confidence distribution for CrowdFlower experiment 1. Each column shows the distribution in percentages for confidence of annotating antecedents of that shell noun. The final row shows the average confidence of the distribution. Number of ASN instances = 2,822. *F* = *fact*, *R* = *reason*, *I* = *issue*, *D* = *decision*, *Q* = *question*, *P* = *possibility*.

One way to measure the reliability of the data, without taking chance correction into account, is to consider the distribution of the ASN instances with different levels of CrowdFlower confidence. Table 2 shows the percentages of instances in different confidence level bands for each shell noun as well as for all instances. For example, for the shell noun *fact*, 8% of the total number of *this fact* instances were annotated with $c < 0.5$. As we can see, most of the instances of the shell nouns *fact*, *reason*, *question*, and *possibility* were annotated with high confidence. In addition, most of them occurred in the band $0.8 \leq c \leq 1$. There are relatively few instances with low confidence for these nouns, suggesting the feasibility of reliable antecedent annotation for these nouns. By contrast, the mental nouns *issue* and *decision* had a large number of low-confidence ($c < 0.5$) instances, bringing in the question of reliability of antecedent annotation of these nouns.

Given these results with different confidence levels, the primary question is what confidence level should be considered acceptable? For our task, we required that at least four trusted annotators out of eight annotators should agree on an answer for it to be acceptable.[9] We will talk about acceptability later in Section 7.

6.2 CrowdFlower experiment 2

Recall that this experiment was about identifying the precise antecedent text segment given the sentence containing the antecedent. It is not clear what the best way to measure the amount of such

[9]We chose this threshold after systematically examining instances with different confidence levels.

	Jaccard			Dice		
	D_o	D_e	α	D_o	D_e	α
A&P	.53	.95	.45	.43	.94	.55
Our results	.47	.96	.51	.36	.92	.61

Table 3: Agreement using Krippendorff's α for CrowdFlower experiment 2. A&P = Artstein and Poesio (2006).

	F	*R*	*I*	*D*	*Q*	*P*	*all*
$c < .5$	11	17	32	31	14	28	21
$.5 \leq c < .6$	12	12	19	23	9	19	15
$.6 \leq c < .8$	36	33	34	32	30	36	33
$.8 \leq c < 1.$	24	22	10	10	21	13	18
$c = 1.$	17	16	5	3	26	4	13
Average c	.74	.71	.60	.59	.77	.62	.68

Table 4: CrowdFlower confidence distribution for CrowdFlower experiment 2. Each column shows the distribution in percentages for confidence of annotating antecedents of that shell noun. The final row shows the average confidence of the distribution. Number of ASN instances = 2,323. *F* = *fact*, *R* = *reason*, *I* = *issue*, *D* = *decision*, *Q* = *question*, *P* = *possibility*.

agreement is. Agreement coefficients such as Cohen's κ underestimate the degree of agreement for such annotation, suggesting disagreement even between two very similar annotated units (e.g., two text segments that differ in just a word or two). We present the agreement results in three different ways: Krippendorff's α with distance metrics Jaccard and Dice (Artstein and Poesio, 2006), Krippendorff's unitizing alpha (Krippendorff, 2013), and CrowdFlower confidence values.

Krippendorff's α using Jaccard and Dice To compare our agreement results with previous efforts to annotate such antecedents, following Artstein and Poesio (2006), we computed Krippendorff's α using distance metrics Jaccard and Dice. The general form of coefficient α is:

$$\alpha = 1 - \frac{D_o}{D_e} \quad (1)$$

where D_o and D_e are observed and expected disagreements respectively. $\alpha = 1$ indicates perfect reliability and ${}_u\alpha = 0$ indicates the absence of reliability. When ${}_u\alpha < 0$, either the sample size is very small or the disagreement is systematic. Table 3 shows the agreement results. Our agreement results are comparable to Artstein and Poesio's agreement results. They had 20 annotators annotating 16 anaphor instances with segment antecedents, whereas we had 8 annotators annotating 2,323 ASN instances. As Artstein and Poesio point out, expected disagreement in case of such antecedent annotation is close to maximal, as there is little overlap between segment antecedents of different anaphors and therefore α pretty much reflects the observed agreement.

Krippendorff's *unitizing* α (${}_u\alpha$) Following Kolhatkar and Hirst (2012), we use ${}_u\alpha$ for measuring reliability of the ASN antecedent annotation task. This coefficient is appropriate when the annotators work on the same text, identify the units in the text that are relevant to the given research question, and then label the identified units (Krippendorff, p.c.). The general form of coefficient ${}_u\alpha$ is the same as in equation 1. In our context, the annotators work on the same text, the ASN instances. We define an *elementary annotation unit* (the smallest separately judged unit) to be a word token. The annotators identify and locate ASN antecedents for the given anaphor in terms of sequences of elementary annotation units.

${}_u\alpha$ incorporates the notion of distance between strings by using a distance function which is defined as the square of the distance between the non-overlapping tokens in our case. The distance is 0 when the annotated units are exactly the same, and is the summation of the squares of the unmatched parts if they are different. We compute observed and expected disagreement as explained by Krippendorff (2013, Section 12.4). For our data, ${}_u\alpha$ was 0.54.[10] ${}_u\alpha$ was lower for the mental nouns *issue* and *decision* and the modal noun *possibility* compared to other shell nouns.

CrowdFlower confidence results We also examined different confidence levels for ASN antecedent annotation. Table 4 gives confidence results for all instances and for each noun. In contrast with Table 2, the instances are more evenly distributed here. As in experiment 1, the mental nouns *issue* and *decision* had many low confidence instances. For the modal noun *possibility*, it was easy to identify the sentence containing the antecedent, but pinpointing the precise antecedent

[10]Note that ${}_u\alpha$ reported here is just an approximation of the actual agreement as in our case the annotators chose an option from a set of predefined options instead of marking free spans of text.

turned out to be difficult.

Now we discuss the nature of disagreement in ASN annotation.

Disagreement in experiment 1 There were two primary sources of disagreement in experiment 1. First, the annotators had problems agreeing on the answer *None*. We instructed them to choose *None* when the sentence containing the antecedent was not labelled. Nonetheless, some annotators chose sentences that did not precisely contain the actual antecedent but just hinted at it. Second, sometimes it was hard to identify the precise antecedent sentence as the antecedent was either present in the blend of all labelled sentences or there were multiple possible answers, as shown in example (2).

(2) Any biography of Thomas More has to answer one fundamental question. Why? Why, out of all the many ambitious politicians of early Tudor England, did only one refuse to acquiesce to a simple piece of religious and political opportunism? What was it about More that set him apart and doomed him to a spectacularly avoidable execution?

The innovation of Peter Ackroyd's new biography of More is that he places the answer to **this question** outside of More himself.

Here, the author formulates the question in a number of ways and any question mentioned in the preceding text can serve as the antecedent of the anaphor *this question*.

Hard instances Low agreement can indicate different problems: unclear guidelines, poor-quality annotators, or difficult instances (e.g., not well understood linguistic phenomena) (Artstein and Poesio, 2006). We can rule out the possibility of poor-quality annotators for two reasons. First, we consider 8 diverse annotators who work independently. Second, we use CrowdFlower's quality-control mechanisms and hence allow only trustworthy annotators to annotate our texts. Regarding instructions, we take inter-annotator agreement as a measure for feasibility of the task, and hence we keep the annotation instruction as simple as possible. This could be a source of low agreement. The third possibility is hard instances. Our results show that the mental nouns *issue* and *decision* had many low-confidence instances, suggesting the difficulty associated with the interpretation of these nouns (e.g., the very idea of what counts as an issue is fuzzy). The shell noun *decision* was harder because most of its instances were court-decision related articles, which were in general hard to understand.

Different strings representing similar concepts As noted in Section 4, the main challenge with the ASN annotation task is that different antecedent candidates might represent the same concept and it is not trivial to incorporate this idea in the annotation process. When five trusted annotators identify the antecedent as *but X* and three trusted annotators identify it as merely *X*, since CrowdFlower will consider these two answers to be two completely different answers, it will give the answer *but X* a confidence of only about 0.6. ${}_u\alpha$ or α with Jaccard and Dice will not consider this as a complete disagreement; however, the coefficients will register it as a difference. In other words, the difference functions used with these coefficients do not respond to semantics, paraphrases, and other similarities that humans might judge as inconsequential. One way to deal with this problem would be clustering the options that reflect essentially the same concepts before measuring the agreement. Some of these problems could also be avoided by formulating instructions for marking antecedents so that these differences do not occur in the identified antecedents. However, crowdsourcing platforms require annotation guidelines to be clear and minimal, which makes it difficult to control the annotation variations.

7 Evaluation of Crowd Annotation

CrowdFlower experiment 2 resulted in 1,810 ASN instances with $c > 0.5$. The question is how good are these annotations from experts' point of view.

To examine the quality of the crowd annotation we asked two judges A and B to evaluate the *acceptability* of the crowd's answers. The judges were highly-qualified academic editors: A, a researcher in Linguistics and B, a translator with a Ph.D. in History and Philosophy of Science. From the crowd-annotated ASN antecedent data, we randomly selected 300 instances, 50 instances per shell noun. We made sure to choose instances with borderline confidence ($0.5 \leq c < 0.6$), medium confidence ($0.6 \leq c < 0.8$), and high confidence ($0.8 \leq c \leq 1.0$). We asked the judges to rate the acceptability of the crowd-answers based on the extent to which they provided interpretation of the corresponding anaphor. We gave them four options: *perfectly* (the crowd's answer is perfect and the judge would have chosen the same antecedent), *reasonably* (the crowd's answer is acceptable and is close to their answer),

		Judge B				
		P	*R*	*I*	*N*	Total
Judge A	*P*	**171**	44	11	7	233
	R	12	**27**	7	4	50
	I	2	4	**6**	1	13
	N	1	2	0	**1**	4
	Total	186	77	24	13	300

Table 5: Evaluation of ASN antecedent annotation. *P = perfectly*, *R = reasonably*, *I = implicitly*, *N = not at all*

implicitly (the crowd's answer only implicitly contains the actual antecedent), and *not at all* (the crowd's answer is not in any way related to the actual antecedent).[11] Moreover, if they did not mark *perfectly*, we asked them to provide their antecedent string. The two judges worked on the task independently and they were completely unaware of how the annotation data was collected.

Table 5 shows the confusion matrix of the ratings of the two judges. Judge B was stricter than Judge A. Given the nature of the task, it was encouraging that most of the crowd-antecedents were rated as *perfectly* by both judges (72% by A and 62% by B). Note that *perfectly* is rather a strong evaluation for ASN antecedent annotation, considering the nature of ASN antecedents themselves. If we weaken the acceptability criteria and consider the antecedents rated as *reasonably* to be also acceptable antecedents, 84.6% of the total instances were acceptable according to both judges.

Regarding the instances marked *implicitly*, most of the times the crowd's answer was the closest textual string of the judges' answer. So we again might consider instances marked *implicitly* as acceptable answers.

For a very few instances (only about 5%) either of the judges marked *not at all*. This was a positive result and suggests success of different steps of our annotation procedure: identifying broad region, identifying the set of most likely candidates, and identifying precise antecedent. As we can see in Table 5, there were 7 instances where the judge A rated *perfectly* while the judge B rated *not at all*, i.e., completely contradictory judgements. When we looked at these examples, they were rather hard and ambiguous cases. An example is shown in (3). The *whether* clause marked in the preceding sentence is the crowd's answer. One of our judges rated this answer as *perfectly*, while the other rated it as *not at all*. According to her the correct antecedent is *whether Catholics who vote for Mr. Kerry would have to go to confession*.

(3) Several Vatican officials said, however, that any such talk has little meaning because the church does not take sides in elections. But the statements by several American bishops that Catholics who vote for Mr. Kerry would have to go to confession have raised the question in many corners about **whether this is an official church position**.

The church has not addressed **this question** publicly and, in fact, seems reluctant to be dragged into the fight..."

There was no notable relation between the annotator's rating and the confidence level: many instances with borderline confidence were marked *perfectly* or *reasonably*, suggesting that instances with $c \geq 0.5$ were reasonably annotated instances, to be used as training data for ASN resolution.

8 Conclusion

In this paper, we addressed the fundamental question about feasibility of ASN antecedent annotation, which is a necessary step before developing computational approaches to resolve ASNs. We carried out crowdsourcing experiments to get native speaker judgements on ASN antecedents. Our results show that among 8 diverse annotators who worked independently with a minimal set of annotation instructions, usually at least 4 annotators converged on a single ASN antecedent. The result is quite encouraging considering the nature of such antecedents.

We asked two highly-qualified judges to independently examine the quality of a sample of crowd-annotated ASN antecedents. According to both judges, about 95% of the crowd-annotations were acceptable. We plan to use this crowd-annotated data (1,810 instances) as training data for an ASN resolver. We also plan to distribute the annotations at a later date.

Acknowledgements

We thank the CrowdFlower team for their responsiveness and Hans-Jörg Schmid for helpful discussions. This material is based upon work supported by the United States Air Force and the Defense Advanced Research Projects Agency under Contract No. FA8650-09-C-0179, Ontario/Baden-Württemberg Faculty Research Exchange, and the University of Toronto.

[11] Before starting the actual annotation, we carried out a training phase with 30 instances, which gave an opportunity to the judges to ask questions about the task.

References

Ron Artstein and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pages 56–63, Potsdam, Germany.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

Simon Philip Botley. 2006. Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics*, 11(1):73–112.

Donna K. Byron. 2003. Annotation of pronouns and their antecedents: A comparison of two domains. Technical report, University of Rochester. Computer Science Department.

Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2009. Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 57–62, Suntec, Singapore, August. Association for Computational Linguistics.

Bin Chen, Jian Su, Sinno Jialin Pan, and Chew Lim Tan. 2011. A unified event coreference resolution by integrating multiple resolvers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 102–110, Chiang Mai, Thailand, November.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37.

Stefanie Dipper and Heike Zinsmeister. 2011. Annotating abstract anaphora. *Language Resources and Evaluation*, 69:1–16.

Karën Fort, Adeline Nazarenko, and Sophie Rosset. 2012. Modeling the complexity of manual annotation tasks: a grid of analysis. In *24th International Conference on Computational Linguistics*, pages 895–910.

Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. 2009. Play the language: Play coreference. In *Proceedings of the Association of Computational Linguistics and International Joint Conference on Natural Language Processing 2009 Conference Short Papers*, pages 209–212, Suntec, Singapore, August. Association for Computational Linguistics.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Boulder, Colorado, June. Association for Computational Linguistics.

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 1255–1265, Jeju Island, Korea, July. Association for Computational Linguistics.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, third edition.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, Jeju Island, Korea, July. Association for Computational Linguistics.

Nitin Madnani, Jordan Boyd-Graber, and Philip Resnik. 2010. Measuring transitivity using untrained annotators. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 188–194, Los Angeles, June. Association for Computational Linguistics.

Rebecca J. Passonneau. 1989. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 51–59, Vancouver, British Columbia, Canada, June. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proeedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Sameer S. Pradhan, Lance A. Ramshaw, Ralph M. Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the International Conference on Semantic Computing*, pages 446–453, September.

Hans-Jörg Schmid. 2000. *English Abstract Nouns As Conceptual Shells: From Corpus to Cognition*. Topics in English Linguistics 34. De Gruyter Mouton, Berlin.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October. Association for Computational Linguistics.

Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang, and Gabriel Othero. 2002. Coreference and anaphoric relations of

demonstrative noun phrases in multilingual corpus. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Conference (DAARC 2002)*, pages 385–427, Lisbon, Portugal, September.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2012. Perspectives on crowdsourcing annotations for natural language processing. In *Language Resources and Evaluation*, volume in press, pages 1–23. Springer.

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. In *Language and Cognitive Processes*, pages 107–135.