Resolving Discourse-Deictic Pronouns: A Two-Stage Approach to Do *It*

Sujay Kumar Jauhar
Carnegie Mellon University
Pittsburgh, PA 15213, USA
sjauhar@cs.cmu.edu

Raul D. Guerra
University of Maryland
College Park, MD 20742, USA
rguerra@cs.umd.edu

Edgar Gonzàlez
Google Research
Mountain View, CA 94043, USA
edgargip@google.com

Marta Recasens
Google Research
Mountain View, CA 94043, USA
recasens@google.com

Abstract

Discourse deixis is a linguistic phenomenon in which pronouns have verbal or clausal, rather than nominal, antecedents. Studies have estimated that between 5% and 10% of pronouns in non-conversational data are discourse deictic. However, current coreference resolution systems ignore this phenomenon. This paper presents an automatic system for the detection and resolution of discourse-deictic pronouns. We introduce a two-step approach that first recognizes instances of discourse-deictic pronouns, and then resolves them to their verbal antecedent. Both components rely on linguistically motivated features. We evaluate the components in isolation and in combination with two state-of-the-art coreference resolvers. Results show that our system outperforms several baselines, including the only comparable discourse deixis system, and leads to small but statistically significant improvements over the full coreference resolution systems. An error analysis lays bare the need for a less strict evaluation of this task.

1 Introduction

Coreference resolution is a central problem in Natural Language Processing with a broad range of applications such as summarization (Steinberger et al., 2007), textual entailment (Mirkin et al., 2010), information extraction (McCarthy and Lehnert, 1995), and dialogue systems (Strube and Müller, 2003). Traditionally, the resolution of noun phrases (NPs) has been the focus of coreference research (Ng, 2010). However, NPs are not the only participants in coreference, since verbal or clausal mentions can also take part in coreference relations. For example, consider:

(1) The United States says **it** may invite Israeli and Palestinian negotiators to Washington.

(2) Without planning **it** in advance, they chose to settle here.

In (1), the antecedent of the pronoun is an NP, while in (2) the antecedent[1] is a clause[2] (Webber, 1988). Current state-of-the-art coreference resolution systems (Lee et al., 2011; Fernandes et al., 2012; Durrett and Klein, 2014; Björkelund and Kuhn, 2014) focus on the former and ignore the latter cases.

Corpus studies across several languages (Eckert and Strube, 2000; Botley, 2006; Recasens, 2008) have estimated that between 5% and 10% of pronouns in non-conversational data, and up to 20% in conversational, have verbal antecedents. A coreference system that is able to handle discourse deixis will thus be more accurate, and benefit downstream applications.

In this paper we present an automatic system that processes discourse-deictic pronouns. We resolve the three pronouns *it*, *this* and *that*, which can appear in linguistic contexts that reflect the phenomenon illustrated in (2). Our system has a modular architecture consisting of two independent stages: classification and resolution. The first stage classifies a pronoun as discourse deictic (or not), and the second stage resolves discourse-deictic pronouns to verbal antecedents. Both stages use linguistically moti-

[1] Since the pronoun in (2) is cataphoric, it has a *postcedent* rather than an *antecedent*, but we use the two indistinctively.

[2] Following the OntoNotes convention, we represent clausal antecedents by their verbal head.

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 299–308,
Denver, Colorado, June 4–5, 2015.

vated features.

We first evaluate our system by measuring the performance of the detection and resolution components in isolation. They outperform several baselines, including Müller's (2007) approach, which is the only other comparable discourse deixis system, to the best of our knowledge. We also measure the impact of our system on two state-of-the-art coreference resolution systems (Durrett and Klein, 2014; Björkelund and Kuhn, 2014). The results show the benefits of stacking a discourse deixis engine on top of NP coreference resolution.

Algorithm 1
Discourse deixis resolution of pronoun p

$p_c(p) \leftarrow \Theta_c(p)$ ▷ Classify
if $p_c(p) > th_c$ **then**
  for $v \leftarrow Candidates(p)$ **do**
    $p_r(v, p) = \Theta_r(v, p)$ ▷ Resolve
  end for
  $v_{best} \leftarrow \arg\max_v p_r(v, p)$
  if $p_r(v_{best}, p) > th_r$ **then**
    return v_{best}
  end if
end if
return $\emptyset$ ▷ No verbal antecedent

2 Related Work

Coreference resolution systems mostly focus on NPs. Although some isolated efforts have been made to study discourse-deictic pronouns, they consist mostly of theoretical inquiries or corpus analyses. A few practical implementations have been proposed as well, but most rely on manual intervention or only apply to restricted domains.

Webber (1988) presents a seminal account of discourse-deictic pronouns. She catalogs how the usage of certain pronouns varies based on discourse context. She also provides an insight into the distinguishing characteristics of discourse deixis.

Several empirical studies have also been conducted to evaluate the prevalence of discourse deixis in corpora across languages. These have been applied to English for dialogues (Byron and Allen, 1998; Eckert and Strube, 2000) and news and literature (Botley, 2006), Danish and Italian (Navarretta and Olsen, 2008; Poesio and Artstein, 2008; Caselli and Prodanof, 2010), and Spanish (Recasens, 2008). These studies find that discourse deixis occurs in different languages, although prevalence depends on the domain in question. While discourse deixis can account for up to 20% of pronouns in dialogue and conversational text, a more general figure is between 5% to 10% for other genres.

In addition to a corpus analysis, Eckert and Strube (2000) provide a schema for performing discourse deixis resolution that they evaluate by measuring inter-annotator agreement on five dialogues from the Switchboard corpus. Byron (2002) presents an early attempt at a practical system that handles discourse deixis. However, it relies on sophisticated discourse and semantic features, thus only working with manual intervention in a limited domain.

The first fully automatic system to handle discourse-deictic pronouns was the one by Müller (2007). In contrast to our two-stage approach, it directly resolves pronouns to nominal or verbal antecedents. The author targets coreference resolution in dialogues, but includes several features that are equally applicable to text data—thus making a comparison to our system viable.

Chen et al. (2011) present another unified approach to dealing with entity and event coreference. Their system combines the predictions from seven distinct mention-pair resolvers, each of which focuses on a specific pair of mention types (NP, pronoun, verb). In particular, their verb-pronoun resolver falls within the scope of discourse deixis. Due to the tight coupling of multiple resolvers, a direct comparison with systems focusing on discourse deixis is hard. However, their features are among the ones considered in this work.

3 Our Approach

In this section we describe the architecture of our two-stage system, and then detail the features used in both stages.

3.1 System Architecture

We propose a two-stage approach for discourse deixis processing. Our system first classifies a potential pronoun as discourse deictic (or not), and then it optionally resolves discourse-deictic pronouns with their antecedent.

Feature	Description	Cla.	Res.	Mül.
Pronoun word	Word of p	•	-	
Demonstrative	p is *this* or *that*	•	-	•
Token position	Relative position of p in sentence	•		
Document position	Relative position of sentence containing p	-		
Verb presence	Sentences before p have verb	•		
Parent lemma	Lemma of parent of p if verb	•		
Parent & label	Lemma of parent and dependency label of p	•	•	
Tree depth	Depth of p in parse tree	-		
Pronoun path	Dependency label path of p to root	•	-	
⋆Negated parent	Parent of p is a negated verb	-		
⋆Parent transitivity	Transitivity of parent verb of p	•		
⋆Clause-governing parent	Probability of parent verb to govern a clause	•		
⋆Attribute lemma	Lemma of attribute of p	-		
⋆Attribute POS	POS of attribute of p	-		
Sentence distance	Number of sentences between v and p		•	•
Token distance	Log-distance between v and p in tokens		•	•
Verb distance	Number of verbs between v and p		-	
Relative position	v precedes p (anaphora/cataphora)		•	
Direct dominance	v is the immediate parent of p		•	
Dominance	v is an ancestor of p		•	•
Candidate path	Dependency label path of v to root		•	
⋆Negated candidate	v is negated		•	
⋆Candidate transitivity	Transitivity of v		•	•
⋆Clause-governing candidate	Probability of v to govern a clause		-	
⋆Right frontier	v is in the right frontier of p		•	•
⋆I-incompatibility	Attribute of p is a non-individual adjective		•	•
⋆Verb association strength	NPMI between v and parent verb of p		-	
⋆Selectional preference	Preference between v and parent verb of p		-	

Table 1: Features used for pronoun p and candidate v in the classification (Cla.) and resolution (Res.) stages. Features marked with • were selected, and those marked with - were discarded by feature selection. The last column (Mül.) contains the features used by Müller (2007). Features marked with ⋆ are described in Section 3.2.

More specifically, and as described in Algorithm 1, a classification model Θ_c is applied to each pronoun p to obtain its probability of being discourse deictic $p_c(p)$. If the probability is above a threshold th_c, the pronoun is considered for resolution. All verbs v in the current and n previous sentences[3] are considered as candidates. A resolution model Θ_r is applied to each candidate v to obtain its probability of being the antecedent of p, $p_r(v, p)$; if the candidate with the highest score v_{best} is above a threshold th_r, then it is returned as the antecedent. Otherwise, the pronoun remains unlinked.

Both components are implemented as maximum entropy classifiers. For simplicity, our approach is independent from the NP–NP coreference resolution component: competition between verbal and nominal antecedents is not considered.

3.2 Features

Table 1 gives an overview of the features that were used by the classification and resolution models. We consider all the features listed in the table, but some of them (marked with -) are pruned by feature selection (see Section 4.2). Real-valued features are

[3] A window of 3 sentences is used in our experiments.

quantized, and dependency label paths are considered up to length 2. Details for the more sophisticated features (marked with * in the table) follow.

Negated parent/candidate We consider a verb token to be *negated* if it has a child connected with a negation label.

Parent/candidate transitivity We consider a verb token to be *transitive* if it has a child with a direct object label.

Clause-governing parent/candidate This is the probability of the parent/candidate to have a clausal or verbal argument. Probabilities for every verbal lemma are estimated from the Google News corpus. We then use the logarithm of these probabilities as the feature values.

Attribute lemma/POS If the pronoun is the subject of a copular verb, we consider the lemma and POS of the attribute of this verb as features.

Right frontier Webber (1988) proposes the *right frontier* condition to restrict the set of candidates available as antecedents for discourse-deictic pronouns. We define this condition in terms of what Webber calls *discourse units*. These are minimal discourse segments, and a sequence of several units can also be nested and form a larger unit. She states that only units on the right frontier (i.e., not followed by another unit at the same nesting level) can be antecedents for such pronouns.

(3) [President Obama *arrived* in San Francisco on Sunday.] [[When he *held* a press conference,] he reported [he would meet with business leaders.]] [He thought **it** went well.]

In (3), where discourse units are marked by square brackets, the verbal heads of discourse segments that are on the right frontier are underlined, while the others are italicized to denote inaccessibility.

In our system, we approximate discourse units by sentences and clauses. The candidate antecedents are the respective verbal heads of these units. This feature triggers if the antecedent candidate occurs on the right frontier of the pronoun. Since we also consider cataphoric relations, we reverse the rule to check the left frontier for these cases.

I-incompatibility Eckert and Strube (2000) define an anaphor to be *I-incompatible* if it occurs in a context in which it "cannot refer to an individual object." Adjectives can be used as contextual cues for I-incompatible anaphors in copular constructions (4).

(4) **It** is true.

Similarly to Müller (2007), we define the *I-incompatibility score* of an adjective as its conditional probability of being the attribute of a non-nominal subject given that it occurs in a copular construction. This is estimated from the Google News corpus as the number of occurrences of the adjective in one of these patterns:

- `clausal subject + BE + ADJ` (*To read is healthy*)
- `IT + BE + ADJ + TO/THAT` (*It is healthy to read*)
- `nominalized`[4] `subject + BE + ADJ` (*The construction was suspended*)
- `-ing subject + BE + ADJ` (*Reading is healthy*)

divided by its number of occurrences in the pattern `BE + ADJ`. At classification time, if the pronoun is in a copular construction with an adjective attribute, the I-incompatibility score of the latter is used as feature.

Verb association strength To capture the strength of association between the candidate antecedent and the parent verb of the pronoun, we use the normalized pointwise mutual information of the two verbs co-occurring within a window of 3 sentences, estimated from counts in the Google News corpus.

Selectional preference We use selectional preference, as defined by Resnik (1997), to capture the degree to which the antecedent makes a reasonable substitute of the pronoun in the context of its parent verb. The selectional preference strength of verb ω is defined as $S_R(\omega) = KL(p(a|\omega) \parallel p(a))$, where KL denotes Kullback-Leibler divergence and a are all possible arguments of ω in the Google News corpus. Larger values of

[4]Nominalizations were identified using NOMLEX (Macleod et al., 1998).

Pronoun	Total	Discourse-Deictic
it	1310	75
that	400	120
this	365	57
Overall	2075	252

Table 2: Distribution of discourse-deictic pronouns in the test set of the CoNLL-2012 English corpus.

this quantity correspond to more selective predicates. Then, the selectional preference strength of a verb ω for a particular argument a is defined as $A_R(\omega, a) = p(a|\omega) \cdot \log\left(p(a|\omega)/p(a)\right)/S_R(\omega)$. To account for nominalizations, verbs and nouns are stemmed following Porter (1980).

4 Evaluation

In this section we describe the setup for evaluating our system.

4.1 Dataset

We perform all our experiments on the English section of the CoNLL-2012 corpus (Pradhan et al., 2012), which is based on OntoNotes (Pradhan et al., 2007). It consists of 2384 documents (1.6M words) from a variety of domains: news, broadcast conversation, weblogs, etc. It is annotated with POS tags, syntax trees, word sense annotation, coreference relations, etc. The coreference layer includes verbal mentions.

Given these annotations, we consider a pronoun to be discourse deictic if the preceding mention in its coreference cluster is verbal, or if it is the first mention in the cluster and the next one is verbal. The distribution of potentially discourse-deictic pronouns (*it*, *this* and *that*) in the test set is summarized in Table 2.

For all our experiments we train, tune and test according to the CoNLL-2012 split of OntoNotes. The gold analyses provided for the shared task are used for training, and the system analyses for development and testing.

4.2 Experiments

We train the two components of our system separately. For each of them, a maximum entropy model is learned on the train partition. Feature selection and threshold tuning are performed by hill climbing on the development set. We use separate thresholds for *it*, *this*, and *that*, since their distributions in the corpus are quite different.

We perform two evaluations of our system: first classification and resolution are evaluated in isolation, and then both components are stacked on top of an NP coreference engine.

For classification, we measure system performance on standard precision (P), recall (R) and F1 of correctly predicting whether a pronoun is discourse deictic or not. For resolution, precision is computed as the fraction of predicted antecedents that are correct, and recall as the fraction of gold antecedents that are correctly predicted. To decouple the evaluation of both stages, we also include results with oracle classifications as input to the resolution stage.

Finally, we use the output of our system to extend the predictions of two state-of-the-art NP coreference systems:

- BERKELEY (Durrett and Klein, 2014), a joint model for coreference resolution, named entity recognition, and entity linking.
- HOTCOREF (Björkelund and Kuhn, 2014), a latent-antecedent model which exploits nonlocal features via beam search.

We only add our predictions for pronouns *it, this, that* that are output as singletons by the NP coreference system.

We report the standard coreference measures on the combined outputs using the updated CoNLL scorer v7 (Pradhan et al., 2014). Here, the systems are evaluated on all nominal, pronominal, and verbal mentions. The metrics include precision, recall and F1 for MUC, B^3 and $CEAF_e$, and the CoNLL metric, which is the arithmetic mean of the first three F1 scores.

4.3 Baselines

We compare our classification component against two baselines:

- ALL, which blindly classifies all mentions as discourse deictic.
- $NAIVE_c$, which classifies all *this* and *that* mentions as discourse deictic, and all *it* mentions as non-discourse-deictic. This is motivated by

	it			*that*			*this*			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ALL	5.7	**100.0**	**10.8**	30.0	**100.0**	46.2	15.6	**100.0**	27.0	12.1	**100.0**	21.7
NAIVE$_c$	0.0	0.0	0.0	30.0	**100.0**	46.2	15.6	**100.0**	27.0	23.1	70.2	34.8
TWOSTAGE	**33.3**	4.0	7.1	**33.6**	77.5	**46.9**	**57.1**	21.1	**30.8**	**35.2**	42.9	**38.6**

Table 3: Classification evaluation (TWOSTAGE corresponds to our system).

	it			*that*			*this*			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
NAIVE$_r$	30.7	30.7	30.7	47.5	**47.5**	47.5	33.3	33.3	33.3	39.3	39.3	39.3
MÜLLER$_r$	30.7	30.7	30.7	47.8	45.0	46.4	43.9	43.9	43.9	41.6	40.5	41.0
TWOSTAGE	**46.3**	**33.3**	**38.8**	**59.6**	46.7	**52.3**	**59.1**	**45.6**	**51.5**	**55.7**	**42.5**	**48.2**

Table 4: Resolution evaluation with oracle classification (TWOSTAGE corresponds to our system).

	it			*that*			*this*			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
NAIVE$_r$	0.0	0.0	0.0	15.3	34.2	21.1	20.0	7.0	10.4	15.3	17.9	16.5
MÜLLER$_r$	0.0	0.0	0.0	16.7	36.7	22.9	20.0	7.0	10.4	16.5	19.0	17.7
TWOSTAGE	**14.3**	**1.3**	**2.4**	**21.5**	**40.0**	**28.0**	**46.2**	**10.5**	**17.1**	**22.6**	**21.8**	**22.2**

Table 5: Resolution evaluation with system classification (TWOSTAGE corresponds to our system).

the distribution of discourse deixis in the corpus (see Table 2).

For resolution, we use the baselines:

- NAIVE$_r$, which resolves a pronoun to the closest verb in the previous sentence. This is motivated by corpus analyses studying the position of discourse-deictic pronouns relative to their antecedents (Navarretta, 2011).
- MÜLLER$_r$, which is an equivalent maximum entropy model using the subset of our features also considered by Müller (2007). See column *Mül.* in Table 1.

Finally, when measuring the impact of our system on top of an NP coreference resolution engine, we consider the following baselines:

- NAIVE, which uses NAIVE$_c$ and NAIVE$_r$.
- MÜLLER, which does not include a classification stage, and uses MÜLLER$_r$ for resolution.
- ONESTAGE, which does not include a classification stage, and uses our complete feature set for resolution.[5]
- ORACLE, which outputs the gold annotations for discourse-deictic relations.

5 Results

The results for the classification stage are presented in Table 3, broken down by pronoun type. ALL performs the poorest overall, penalized by a precision just above 12%. Since in the case of *it* only 5.7% of the occurrences are discourse deictic, NAIVE$_c$ gets better results overall by always classifying *it* as non-deictic. Our TWOSTAGE system improves over NAIVE$_c$ by an additional 4% F1. However, the scores remain low—partly because of the difficulty of the problem (especially the class imbalance), and partly because despite using a rich set of features, most of them focus on local context and ignore cues at the discourse level. The classification of *it* is particularly difficult, reflecting the fact that the pronoun has a wide variety of usages in English.

[5]Feature selection and threshold tuning were done separately for this model. The exact subset of resolution features that were chosen is omitted for brevity.

The scores for resolution are shown in Tables 4 and 5. The former uses oracle classification whereas the latter uses the system output of our classifier.

With oracle classification, NAIVE$_r$ and MÜLLER$_r$ perform very similar, except for the case of *this*. Our TWOSTAGE resolver outperforms both of them for all pronouns and metrics, except for the recall of *that*. Overall, the difference in F1 is 9 points over NAIVE$_r$ and 7 points over MÜLLER$_r$. The evaluation actually penalizes recall for our system, since we do not take advantage of the fact that all considered pronouns are discourse deictic: we trust the threshold and do not force the assignment of an antecedent.

All the results are lower with system classification. Given that our classifier performs the best for *that*, the drop for this pronoun is not as high as for the other two. Again, *it* stands out as the hardest pronoun to resolve. Neither NAIVE$_r$ nor MÜLLER$_r$ recover any correct antecedent for *it*. TWOSTAGE obtains the highest scores across all pronouns and metrics.

Finally, Table 6 contains the coreference measures for end-to-end evaluation on top of the BERKELEY and HOTCOREF systems. The ORACLE row shows an upper bound of 2% in CoNLL score improvement. All three baselines—NAIVE, MÜLLER and ONESTAGE—actually cause a decrease of up to 0.9% CoNLL.

Our system TWOSTAGE achieves a small fraction of the headroom. The total number of discourse-deictic entities that it predicts on the test set is 248, of which 204 end up merged in the BERKELEY output, and 210 in HOTCOREF. This allows it to obtain the best B^3, CEAF$_e$ and CoNLL values, despite the fact that the low recall in the classification of discourse-deictic *it* reduces our margin for recall gains by one third. The drop in MUC highlights the difficulty of keeping the precision level, but our system is able to reach a better precision-recall balance than the other compared approaches.

We assess the statistical significance of the improvements of TWOSTAGE over BERKELEY and HOTCOREF using paired bootstrap resampling (Koehn, 2004) followed by two-tailed Wilcoxon signed-rank tests. All the differences are significant at the 1% level, except for the B^3 F1 differences.

Error type	**%**
System errors	
Classification	22.9
Resolution	20.0
Preprocessing	5.7
Annotation errors	
Missing	11.4
Multiple antecedents	20.0
System & Annotation errors	
Debatable	20.0
Overall	100.0

Table 7: Distribution of errors.

6 Error Analysis

In order to gain insight into the precision errors of our system, we manually analyzed 50 of its decisions on the CoNLL-2012 development set. Of these, 30% were correct, matching the gold annotation, as in (5).[6]

(5) Ah, we have *established* the year 2006 as Discover Hong Kong Year. Why is **that**?

The distribution of errors for the remaining cases is shown in Table 7. While half of the errors are due to actual errors in the model learned by our system—either in classification (6) or resolution (7)—or due to a pre-processing error, another third of them are not true errors but missing (8) or partial annotations (9)–(10) in the gold standard corpus.

(6) If pictures are taken without permission, **that** is to say, it will at all times be pursued by legal action, a big hassle.

(7) Do we even *know* if these two medications are going to be effective against a strain that hasn't even presented itself? Here's the important thing about **that**.

(8) You will be taken to stand before governors and kings. People will do **this** to you because you follow me.

[6]The pronoun to be resolved is in boldface, the antecedent annotated in the gold standard (if any) is in italics, and the antecedent predicted by our system is underlined.

		MUC			B^3			$CEAF_e$			CoNLL
		P	R	F1	P	R	F1	P	R	F1	F1
BERKELEY	Durrett and Klein (2014)	**72.61**	69.91	**71.23**	**61.18**	56.43	**58.71**	**56.16**	54.23	55.18	61.71
	+ NAIVE	70.10	**70.33**	70.21	58.64	**57.49**	58.06	52.02	**57.21**	54.50	60.92
	+ MÜLLER	71.57	70.18	70.86	60.15	57.02	58.54	54.55	55.86	55.20	61.53
	+ ONESTAGE	71.63	70.19	70.90	60.21	57.03	58.58	54.66	55.88	55.26	61.58
	+ TWOSTAGE	71.87	70.19	71.02	60.50	57.02	**58.71**	55.14	55.77	**55.45**	**61.73**
	+ ORACLE	*73.09*	*71.64*	*72.36*	*61.95*	*58.77*	*60.32*	*58.05*	*58.51*	*58.28*	*63.65*
HOTCOREF	Björkelund and Kuhn (2014)	**74.30**	67.46	**70.72**	**62.71**	54.96	58.58	**59.40**	52.27	55.61	61.64
	+ NAIVE	71.38	**67.92**	69.61	59.72	**56.09**	57.85	54.14	**55.45**	54.79	60.75
	+ MÜLLER	73.11	67.74	70.32	61.51	55.58	58.39	57.32	54.00	55.61	61.44
	+ ONESTAGE	73.15	67.79	70.37	61.54	55.61	58.43	57.35	54.02	55.64	61.48
	+ TWOSTAGE	73.49	67.77	70.51	61.94	55.58	**58.59**	58.14	53.93	**55.96**	**61.69**
	+ ORACLE	*74.79*	*69.20*	*71.88*	*63.59*	*57.33*	*60.30*	*61.33*	*56.87*	*59.02*	*63.73*

Table 6: End-to-end coreference resolution evaluation (TWOSTAGE corresponds to our system). All differences between the baseline system and TWOSTAGE are significant at the 1% level except for the B^3 F1 differences.

(9) At this point they've *wittled* it down to one aircraft and a missing crew of four individuals. So we've gone from several possible aircraft to one aircraft and from several missing airmen to four. So how much easier will **that** make it for you to unlock this case, do you think?

(10) What do you mean by *that*? Either she either passed out regurgitated. Something had happened. And on top of **that** now there's a statement...

The examples (8)–(10) show the difficulty of annotating discourse deixis relations under guidelines that require a unique verbal antecedent (Poesio and Artstein, 2008; Recasens, 2008). In our analysis we found several cases in which more than one antecedent is acceptable. This is usually the case when there is an elaboration (i.e., both the first clause and the follow-up clause restating or elaborating on the first one are acceptable antecedents, as in (9)) or a sequence of related and overlapping events. As pointed out by Poesio and Artstein (2008), "it is not completely clear the extent to which humans agree on the interpretation of such expressions," and the inconsistencies observed in the data are evidence of this.

Another class of hard cases are the discourse-deictic pronouns that are used for *packaging* a previous fragment or set of clauses (10). It is very hard to pick an antecedent for them, even deciding whether the antecedent is an NP or a clause (Francis, 1994).

Finally, in 20% of the cases the system and the annotation are in disagreement, but both decisions are debatable. In many of them, the system did not make any prediction, but the one in the gold annotation is incorrect. In (11), *act* is a more plausible antecedent for **that**.

(11) "Why didn't the Bank Board act sooner?" he *said*. "**That** is what Common Cause should ask be investigated."

As a result, even though our system obviously makes multiple mistakes in its decisions, we believe that the evaluation overpenalizes its performance due to the debatable and not always clear-cut annotations discussed above. Discourse deixis resolution is a hard problem in itself (the chances of selecting a wrong antecedent for a pronoun are many times greater than picking the right one), and this difficulty is accentuated by the problematic annotations in the training and test data.

Given the difficulty of identifying a single antecedent to discourse-deictic pronouns, as evidenced by the low inter-annotator agreement on this task, a more flexible evaluation measure for discourse deixis systems is needed.

7 Conclusion

We have presented an automatic system for discourse deixis resolution. The system works in two stages: first classifying pronouns as discourse deictic or not, and then assigning an antecedent.

Empirical evaluations show that our system outperforms naive baselines as well as the only existing comparable system. Additionally, when stacked on top of two different state-of-the-art NP coreference resolvers, our system yields improvements on the B^3, $CEAF_e$ and CoNLL measures. The results are still far from the upper bound achievable by an oracle. However, our research highlights the inconsistencies in the annotation of discourse deixis in existing resources, and thus the performance of our system is likely underestimated.

These inconsistencies call for future work to improve existing annotated corpora so that similar systems may be more fairly evaluated. Regarding our approach, a tighter integration between the NP and discourse deixis components could help them make more informed decisions. Other future research includes jointly learning the classification and resolution stages of our system, and exploring semi-supervised learning techniques to compensate for the paucity of annotated data. Finally, we would like to transfer our system to other languages.

Acknowledgments

This work was done when the first two authors were interns at Google. We would like to thank Greg Durrett and Anders Björkelund for kindly sharing the outputs of their systems. Thanks also go to Rajhans Samdani and the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of ACL*, pages 47–57.

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 and James F Allen. 1998. Resolving demonstrative anaphora in the TRAINS93 corpus. In *Proceedings of the 2nd Colloquium on Discourse, Anaphora and Reference Resolution*.

Donna K Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of ACL*, pages 80–87.

Tommaso Caselli and Irina Prodanof. 2010. Annotating event anaphora: A case study. In *Proceedings of LREC*, pages 723–728.

Bin Chen, Jian Su, Sinno Jialin Pan, and Chew Lim Tan. 2011. A unified event coreference resolution by integrating multiple resolvers. In *Proceedings of IJCNLP*, pages 102–110.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Proceedings of CoNLL: Shared Task*, pages 41–48.

Gill Francis. 1994. Labelling discourse: An aspect of nominal-group lexical cohesion. In M. Coulthard, editor, *Advances in Written Text Analysis*, pages 83–101. Routledge, London.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. In *Proceedings of CoNLL: Shared Task*, pages 28–34.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. NOMLEX: A lexicon of nominalizations. In *Proceedings of EURALEX*, pages 187–193.

Joseph F. McCarthy and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of IJCAI*, pages 1060–1065.

Shachar Mirkin, Jonathan Berant, Ido Dagan, and Eyal Shnarch. 2010. Recognising entailment within discourse. In *Proceedings of COLING*, pages 770–778.

Christoph Müller. 2007. Resolving *it*, *this*, and *that* in unrestricted multi-party dialog. In *Proceedings of ACL*, pages 816–823.

Costanza Navarretta and Sussi Olsen. 2008. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC*, pages 2046–2052.

Costanza Navarretta. 2011. Antecedent and referent types of abstract pronominal anaphora. In *Proceedings of the Workshop Beyond Semantics: Corpus-based investigations of pragmatic and discourse phenomena*, pages 99–10.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of ACL*, pages 1396–1411.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC*, pages 1170–1174.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of ICSC*, pages 446–453.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of CoNLL: Shared Task*, pages 1–40.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of ACL*, pages 30–35.

Marta Recasens. 2008. Discourse deixis and coreference: Evidence from AnCora. In *Proceedings of the 2nd Workshop on Anaphora Resolution (WAR II)*, pages 73–82.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57.

Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek. 2007. Two uses of anaphora resolution in summarization. *Information Processing & Management*, 43(6):1663–1680.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL*, pages 168–175.

Bonnie L Webber. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of ACL*, pages 113–122.