Examining the Impact of Coreference Resolution on Quote Attribution

Tim O'Keefe tokeefe@it.usyd.edu.au
Kellie Webster kweb3773@uni.sydney.edu.au
James R. Curran james@it.usyd.edu.au

ə-lab, School of Information Technologies
University of Sydney
NSW 2006, Australia

Abstract

Quote attribution is the task of identifying the speaker of each quote within a document. While recent research has established large-scale corpora for this task, these corpora are not yet consistent in the way they handle candidate speakers, and many of the reported results rely on gold standard annotations of both entities and coreference chains.

In this work we evaluate three quote attribution systems with automatically produced candidate speakers and coreference chains. We perform these experiments over four separate corpora, which allows us to determine how coreference resolution effects quote attribution, and to use the task as an extrinsic evaluation of three coreference systems.

1 Introduction

News articles are often driven by the quotes that appear within them. Approximately 32% of the tokens in the Sydney Morning Herald Corpus (SMHC) (Pareti et al., 2013) appear within a quote. Ignoring the attributed nature of this text can result in incorrectly assigning text to a document's author, rather than to the speaker the author attributes it to. Quote attribution is thus important for applications such as information retrieval, opinion mining, media monitoring, and others.

Early research into quote attribution and quote extraction was largely rule-based, as there was no large-scale data available. Several more recent studies (Elson and McKeown, 2010; O'Keefe et al., 2012; He et al., 2013; Pareti et al., 2013) have addressed this with corpora covering both news articles and literature. However, despite the importance of candidate speakers to this task, work thus far has treated candidates speakers inconsistently. Elson and McKeown (2010) include automatically identified named entities and common nouns, but do not include pronominal references or attempt coreference, which they state is problematic due to the domain (literature). He et al. (2013) include automatically identified named entities with limited gold-standard coreference, but do not include pronouns or common nouns. The SMHC (Pareti et al., 2013) includes gold-standard named entities and pronouns, as well as gold-standard coreference, but does not include common noun candidates. Finally the PARC (Pareti, 2012) is intended to cover attribution more generally, and so does not include any candidate speakers except for those that have attributed text.

Our work addresses the problem of inconsistent candidates within these corpora by separately aligning the output of three coreference resolution systems, Stanford (Lee et al., 2011), Reconcile (Stoyanov et al., 2010), and a naive baseline system, with the gold-standard speaker annotations. We can then evaluate the quote attribution methods from O'Keefe et al. (2012) with a set of speakers that have been identified in a more consistent manner across attribution methods and corpora. O'Keefe et al. note that one of the primary factors confounding their evaluation was that the set of candidates was not consistent, which our work addresses.

Our second main contribution is that we use quote attribution as an extrinsic evaluation for coreference resolution. Intrinsic evaluation of coreference is known to be problematic (Luo, 2005; Stoyanov et al., 2009) and for this reason, Mitkov et al. (2007) proposes extrinsically evaluating it by measuring its impact on downstream processes. Additionally we are able to gauge the impact of coreference resolution on quote attribution in literature, which is a domain that has not been studied in the work on coreference thus far.

Tim O'Keefe, Kellie Webster, James R. Curran and Irena Koprinska. 2013. Examining the Impact of Coreference Resolution on Quote Attribution. In *Proceedings of Australasian Language Technology Association Workshop*, pages 43−52.

2 Background

The first work to use large-scale data and machine learning for this task was the work of Elson and McKeown (2010) (hereafter referred to as EM2010). Their system uses a binary classifier to produce a probability that each of 15 candidates is the speaker, and returns the candidate with the highest probability. For data they build a corpus of direct quotes from 19th century literature, which includes both proper nouns and common nouns as candidate speakers, with the former identified using the Stanford NER system, and the latter identified through their own method. They do not identify pronouns and only perform coreference on the NEs, using a simple system.

Following on from EM2010, was the work of O'Keefe et al. (2012). They note that EM2010 had used some features that relied on gold standard information about previous decisions, which O'Keefe et al. replaced with features using predicted information and a sequence decoding step. They also evaluated their method on two other corpora, one that they build from Sydney Morning Herald[1] news articles (SMHC), and another over Wall Street Journal[2] news articles (PARC) that was introduced in Pareti (2012). They found that removing the gold standard features had a large impact on accuracy, and that their sequence labelling approaches could recover some of that lost accuracy. Later work by Pareti et al. (2013) extended the SMHC to include indirect and mixed quotes, though their focus was on quote extraction.

While the work of O'Keefe et al. (2012) and Pareti et al. (2013) was mainly focused on news articles, He et al. (2013) focused on literature. They developed a model that treated the task similarly to EM2010, though they considered it to be a ranking problem. As part of their work they introduced a new corpus which covers the entirety of the novel *Pride & Prejudice*. While their work outperformed the previous work on literature by EM2010, their system was very slow, so they did not provide a full comparison.

2.1 Coreference resolution

Coreference resolution (e.g. Pradhan et al. (2011)) is the task of partitioning mentions (typically noun phrases) into equivalence classes which refer to the same real world entity. It has largely been framed in terms of anaphoric links; that is, clusters of coreferential mentions are formed by determining whether a particular mention anaphorically points to another preceding it in the text (its antecedent). Both supervised and unsupervised models have been proposed.

The first competitive learning based system is described in Soon et al. (2001). A binary classifier was trained to determine whether pairs of mentions were coreferential, based on 12 features which considered surface level details such as string matching and heuristically determined morphosyntatics. Its feature set was expanded in Ng and Cardie (2002) to include the role of syntactic constraints and modification on coreference. Various works (Bengtson and Roth, 2008; Stoyanov et al., 2010; Stoyanov and Eisner, 2012) have expanded this feature set further.

Ng and Cardie (2002) also proposed ranking potential coreference links. Where Soon et al. assigned the closest positively classified mention as the antecedent of an active mention, ranking approaches define a window for candidate selection and return the most probable candidate within the window. Systems can either incorporate ranking as a post-processing stage which forms clusters based on pairwise probabilities (Ng and Cardie, 2002; Stoyanov et al., 2010; Denis and Baldridge, 2008), or they can rank during clustering (Rahman and Ng, 2009).

Stanford's system (Lee et al., 2011) achieved the best result in the CoNLL 2011 shared task and remained competitive in CoNLL2012 using a simple, unsupervised classifier. It captures global consistency constraints by having cluster level modelling, which it achieves by having a series of sieves that each read the document and expand clusters. The sieves are arranged in order of decreasing precision, such that mentions with a high chance of being coreferential are clustered first, which allows more difficult mentions to use more information from the expanded clusters.

Research into quote attribution has ignored the impact that these different approaches could have, and the four large-scale corpora that exist for quote attribution all include some gold-standard information about either the mentions or the coreference chains. Thus the goal of our work is to use consistent coreference methods across the different corpora, in order to evaluate the effect of coreference on quote attribution. This also allows us to

[1] http://www.smh.com.au

[2] http://www.wsj.com

Corpus		SMHC	PARC	LIT	P&P
Documents		965	2,280	11	1
Tokens		601k	1,139k	407k	144k
Quotations		6,705	9,961	3,486	1,692
Entities	Proper	Gold	Gold	Auto	Auto
	Pronouns	Gold	Gold	-	-
	Common	-	Gold	Auto	-
Coref	Proper	Gold	-	Auto	Gold
	Pronouns	Gold	Gold	-	-
	Common	-	-	-	-

Table 1: Comparison of the four corpora in terms of both size, and the candidate speakers included.

evaluate the coreference methods extrinsically.

3 Corpora

In this work we perform experiments over two corpora containing news articles and two corpora containing works of fiction.

3.1 Sydney Morning Herald Corpus (SMHC)

The original version of the SMHC (O'Keefe et al., 2012) covered all of the direct quotes from 965 articles from the 2009 Sydney Morning Herald. The quotes were extracted automatically, and their speakers were annotated by one of 16 annotators, 11 of whom were employed using the website Freelancer[3], while the remaining 5 were expert annotators. 400 of the documents were double annotated, with raw agreement on the speakers of 98.3%. Later work by Pareti et al. (2013) extended the SMHC by adding indirect and mixed quotes, which was performed by a single annotator.

The candidate speakers for this corpus consist of gold-standard annotations of NEs and pronouns, which were completed as part of a separate research project (Hachey et al., 2013). Both the NEs and the pronouns were manually merged into coreference chains. Annotating a candidate as being the correct speaker of a quote in this corpus involves linking to the *coreference chain*, rather than a specific mention. This corpus does not include any common noun references to entities.

3.2 Penn Attribution Relations Corpus (PARC)

Our next corpus was introduced in Pareti (2011, 2012) and covers 2,280 articles from the Wall Street Journal. Pareti's work includes more general forms of attributable text than we are interested in, so we use just the assertions, as they correspond to quotes. This corpus was built semi-automatically from the Penn Discourse TreeBank (Prasad et al., 2006), which does not include all quotes, so it is not yet fully annotated. Pareti estimates that 30-50% of the corpus is unannotated, which means that there are many articles where legitimate quotes are missed.

As this corpus is not specifically designed for quote attribution, it does not come with any candidate speakers, with the exception of the text that each quote is attributed to. O'Keefe et al. (2012) use the BBN pronoun coreference and entity type corpus (Weischedel and Brunstein, 2005), although with automatically coreferred pronouns. This gives them gold-standard named entities, pronouns, and common nominal references, but only coreference for pronouns. To align Pareti's speakers (called *source*) O'Keefe et al. used the first BBN entity that was a subspan of Pareti's *source* annotations, and where no BBN entity matched, they inserted Pareti's *source* itself as an additional mention. The quotes from Pareti's annotations with an implicit source cannot be automatically linked to any entity, so they were ignored.

3.3 Columbia Quoted Speech Attribution Corpus (LIT)

The LIT corpus was introduced by Elson and McKeown (2010) and constituted the first large-scale corpus of quote attribution. It partially covers 7 short stories and chapters from 4 novels from 19th century fiction. The corpus was annotated using Amazon's Mechanical Turk[4], with three annotations per quote. Disagreements were settled by taking a majority vote, and in their original work, quotes with three-way disagreement were discarded, along with cases of non-dialogue text. Later work (O'Keefe et al., 2012) re-annotated the cases of three-way disagreement and filled in other gaps in the corpus, such that the annotated parts of each text were contiguous.

EM2010 found candidate speakers by identifying NEs with the Stanford NE tagger, and common nouns through patterns looking for a determiner, an optional modifier, and a head noun. They use their own system to link NEs with similar names, though they do not attempt any coreference on

[3] http://www.freelancer.com

[4] http://www.mturk.com/

the common nouns. They do not find pronouns, as they consider coreference to be part of the attribution system's job. In their results over LIT, O'Keefe et al. (2012) did identify pronouns, and used a simple rule-based method to link them to either NEs or common nouns.

3.4 Pride and Prejudice Corpus (P&P)

The final corpus that we use in this work is the corpus introduced by He et al. (2013). This corpus was annotated by an English Major and covers the entirety of the novel *Pride & Prejudice* by Jane Austen. It contains 1,260 quotes, which were extracted automatically.

He et al. also found candidate speakers by using the Stanford NER system, along with a manual preprocessing step where they grouped proper nominal references into sets of aliases for each character. They consider a correct attribution to be from a quote to a character, rather than to a textually-grounded mention of a character. As such, their candidate characters are two proper noun references before and two after each quote, as long as those proper nominal references are within the set of aliases that they manually defined. Since they are trying to explicitly link quotes to characters, they do not consider common or pronominal references as candidates, though they do use them as features. Note that the set of characters that they can attribute quotes to is closed, and does not include any unnamed characters.

3.5 Corpus Comparison

Table 1 shows a comparison of the four corpora. The largest in terms of documents, tokens, and number of quotations is the PARC, although it is worth noting that it is not yet fully annotated. The LIT corpus is also not fully annotated, although as the direct quotations were extracted automatically we know that there are 2,416 quotes that are missing their speakers. The other two corpora (SMHC and P&P) are fully annotated and so give a fair indication of the density of quotes. For this table we only counted quotes where a speaker was given.

In terms of candidate speakers the table shows considerable variance amongst the corpora. All the corpora include proper nominal candidates, although only the SMHC and PARC include gold standard proper nominals. Pronouns and common nominals are more mixed, with only the PARC including gold-standard candidates from these two categories. Coreference information is even less consistent, with the SMHC including gold-standard coreference for the two categories of candidates it contains and P&P including gold-standard coreference for its automatically identified named entities. LIT includes only automatic coreference of named entities, while PARC only includes gold-standard coreference of pronouns.

System	MUC-6		CoNLL-2011	
	MUC	B^3	MUC	B^3
Stanford	*78.2*	*73.8*	59.6	68.9
Reconcile	66.4	70.8	-	-

Table 2: Performance of Stanford and Reconcile on standard test sets using standard evaluation metrics. Results using gold cf. automatically detected mentions are indicated in italics.

4 Coreference Systems

The three coreference resolution systems that we use are Stanford's CoreNLP package (Raghunathan et al., 2010), Reconcile (Stoyanov et al., 2010), and a naive baseline system. By using Stanford and Reconcile we can evaluate the two main types of systems, as they are unsupervised and supervised respectively. The naive system is included for comparison. It performs NE coreference using simple string-matching of NEs found with Stanford's NE tagger, and coreference of pronouns by linking them to the most recent gender-matching antecedent. The naive baseline does not include common noun mentions. We experimented with a fourth system, CherryPicker (Rahman and Ng, 2009), but are unable to include results using CherryPicker as it crashed frequently.

Intrinsic evaluation of coreference resolution is difficult and even the relative performance of different systems can be hard to determine since system performance may be quoted on different corpora, using different evaluation metrics and even in different environments (e.g. the use of gold vs. automatically detected mentions). All of these effects can be seen in Table 2, with results using gold mention boundaries indicated in italics.

In this work we attempt to run all systems with minimal deviation from their default settings. However, since these systems were built for newswire, their architecture is not designed to scale to the longer texts from P&P and LIT, which forced us to make some changes. There were also some further issues that are detailed below.

Stanford

Stanford's mention spans are by design longer than the other two systems, and include overlapping mentions. We greedily kept the smaller mention of any overlapping pair, and retained the non-overlapping fragments from the longer mention as separate mentions. Some fragments and boundary tokens contained extraneous information, which we removed. We also removed the part of any mention following a comma or WH word, so as to retain the head NP. The default setting where all preceding mentions are potential antecedents was kept for the newswire corpora, but for LIT and P&P, a threshold of 100 sentences was used.

Reconcile

Due to memory constraints, the longer of the LIT texts and the training set of P&P were processed in 500 paragraph chunks.

5 Quote Attribution

Given a set of candidate speakers and a quote, quote attribution is the task of determining which of the candidates is the speaker of the quote. We note that for this task it is possible to consider a correct attribution to be either to a textually-grounded mention of an entity (called the *source* of the qoute), or to an entire coreference chain. In many cases the *source* will be a pronoun or common noun, that does not provide much information on its own. Other cases will include no explicit source, such as paragraph-long direct quotes. While our systems consider individual mentions as candidates, we consider a correct attribution to be to a whole coreference chain, meaning that the system can return the wrong textually-grounded mention, but still be considered correct if that mention is clustered with the *source*.

As the focus of this work is on evaluating the impact of coreference resolution on quote attribution, we do not propose any new approaches. Instead we use three of the systems from O'Keefe et al. (2012), namely the rule-based system, a simple binary classifier (called *no sequence* in O'Keefe et al.), and a CRF. All of these systems use the preprocessing described in O'Keefe et al., and all are evaluated using accuracy.

Rule-based

The rule-based system works by returning the candidate speaker that is nearest to either the quote or a speech verb, as long as that candidate is in the paragraph the verb or quote is in, or any preceding it. Speech verbs are identified using the list from Elson and McKeown (2010), and must be found in the same sentence as the quote. If a speech verb is found then the candidate nearest the verb is returned, otherwise it is the candidate nearest the quote. Though this system is very simple, O'Keefe et al. found that it worked about as well as machine learning approaches.

Binary classifier

The binary classifier assigns a binary probability of *speaker* vs. *not speaker* for up to 15 candidate speakers that are mentioned in the paragraph the quote is in or any preceding it. The final decision on which of the 15 candidates is the speaker is made by returning the candidate with the highest *speaker* probability. We use the maximum entropy learner from scikit learn[5]. While this method makes use of machine learning, there is no decoding step to ensure a sensible sequence of speakers, nor is there direct competition for probability mass between candidates. The advantage of this method is that it is able to generate many training instances, as there are effectively up to 15 training instances per quote, rather than the single instance that would be present for a model involving direct competition for probability mass.

Conditional Random Field (CRF)

The final quote attribution method that we use is a CRF which, similarly to the binary classifier, chooses between up to 15 candidate speakers. The difference with the CRF is that it includes a decoding step, and so can forego good local decisions about particular quotes in order to achieve a better sequence of decisions for all of the quotes. It includes a class labelling scheme where the candidates are numbered according to their ordinal position preceding the quote. This labelling scheme forces the candidates to compete for probability mass, although it reduces the number of training instances available to the classifier, and increases the number of features that are considered at each decision point.

6 Speaker Alignment

In order to evaluate the effect of coreference on quote attribution we first align the gold-standard

[5] http://scikit-learn.org

	SMHC			PARC			LIT			P&P		
	Rule	Bin	CRF	Rule	Bin	CRF	Rule	Bin	CRF	Rule	Bin	CRF
Naive	70	73	72	63	65	68	44	46	37	60	62	54
Stanford	68	78	76	80	82	83	40	48	40	44	56	53
Reconcile	69	76	76	77	80	81	37	50	37	45	51	45
Gold	74	78	75	87	92	92	54	54	50	54	62	57

Table 3: Quote attribution results using the source-based alignment method. The gold results use the candidates that come with the corpora.

speaker annotations with the automatically generated coreference chains. These alignment methods erase the gold-standard speaker annotation from each quote and replace it with one of the automatically generated coreference chains, so that the quote attribution methods can learn and predict using predicted coreference chains only. Quotes whose speakers could not be aligned are considered incorrect, as no correct attribution is possible.

Source-based alignment

Not all the corpora have gold-standard coreference chains, so our first alignment method aligns the textually-grounded source of each quote with a mention from the automatic coreference chains. Since speaker predictions are to whole coreference chains, any mention in the automatic coreference chain would then be considered correct. Consider the following example:

> “It doesn’t seem the numbers are there yet, but I will continue to build my case,” Senator Xenophon said.

The textually-grounded source of this quote is ‘Senator Xenophon’, so the source-based alignment works by finding the automatic coreference chain that includes ‘Senator Xenophon’ as a mention. This aligned coreference chain would then be considered the speaker.

While all of the corpora include some annotations of which mention best represents the speaker, there are some individual quotes where these annotations are not included. For these cases we align the automatic coreference chain with the mention from the gold-standard speaker’s coreference chain that is nearest to the quote.

Canonical mention-based alignment

Two of our data sets, SMHC and P&P, include full coreference between the labelled gold-standard mentions, and have annotations of which gold-standard chain represents the speaker. For these two corpora, rather than considering the source of the quote, we use the canonical mention from the speaker’s gold-standard coreference chain. We can then align the canonical mention with a mention from the automatic coreference chains, and again consider any mention from that chain to be the correct speaker. The gold-standard canonical mention will normally be mentioned early in a document, and will be an unambiguous reference to the real-world entity.

	SMHC			P&P		
	Rule	Bin	CRF	Rule	Bin	CRF
Naive	51	54	52	34	47	41
Stan.	40	47	46	32	43	33
Recon.	39	45	43	37	50	38

Table 4: Quote attribution results using the canonical-based alignment method.

7 Results

7.1 Quote attribution

The results in Tables 3 and 4 demonstrate that quote attribution is more successful over news than it is over literature, which agrees with O’Keefe et al. (2012). This is likely due to a number of factors, including the upstream processes being trained over news, the length of the documents, the formality of the text, and that journalists need to clearly identify who is speaking, while authors of fiction have more artistic freedom.

In all but one case the simple binary model outperformed the rule-based approach. This indicates that while the task may appear reasonably straightforward, there is still significant value in using large-scale data to learn a model. In particular some of the gains in literature were as high as 13 percentage points.

While the binary model performed well, the CRF model was somewhat inconsistent. On news text with the source-based alignment method the CRF did nearly as well as and sometimes better than the binary model, and better than the rule-based model. However with the literature text the CRF performed poorly. We found that this was due to some quotes not having a correct speaker within the set of 15 candidates that the learner considered. In these cases the CRF marks the quotes as not having a speaker, however, as these cases tend to cluster together in long dialogue chains in the literature corpora, the CRF learned that it is extremely likely to transition from not having a speaker to not having a speaker. This meant that if the CRF predicted that a single quote had no speaker then it would tend to predict that a number of subsequent quotes had no speaker. By contrast, the rule-based method and the binary model are forced to choose a speaker, and so do not suffer from this problem.

Across all of the corpora the gold-standard results were at least as good, if not better than the results using automatic coreference. This indicates that coreference systems are not over-clustering their results. The most surprising of the gold standard results is on PARC, where they are far better than the automatic results, despite PARC not including full coreference. The reason for this is that the PARC quotes must be attributed to entities within the same sentence as the quote. Both Stanford and Reconcile will tend to produce more mentions than the PARC gold standard, which can confuse the classifier, and Naive will produce no common nominal mentions, so all three automatic systems will perform substantially worse than the gold standard, despite potentially having more coreference information.

7.2 Extrinsic evaluation

Before discussing the results of our extrinsic evaluation, we would first like to note a weakness of our approach. In our framework if any coreference system outputs a single chain containing all mentions, it would score perfectly, as any predicted speaker would be the chain containing the correct mention. While this is not ideal, Vilain et al. (1995)'s MUC F-score has a similar problem, so, as they do, we simply note that this evaluation can not be considered independently of other metrics.

Table 5 shows the number of quotes whose speaker had no corresponding mention in the automatic coreference chains. For the source-based alignment the naive approach had a large number of misses, which is mostly due to the naive system not handling common noun references. This problem is not as severe in the canonical-based alignment, which will in most cases be a proper nominal reference, which the naive method can detect. Interestingly, there were no mentions that could not be aligned in P&P, although it is worth noting that P&P does not include quotes whose speakers are only referenced with common nouns.

	SMHC	PARC	LIT	P&P
Source				
Naive	352	656	214	0
Stanford	19	45	6	0
Reconcile	22	25	13	0
Canonical				
Naive	367			0
Stanford	285			0
Reconcile	297			0

Table 5: Number of gold speakers without a corresponding mention in the automatic coreference chains, for both the source and canonical-based alignment methods.

For the source-based alignment results in Table 3, we note that in almost all cases the coreference systems were able to help the quote attribution systems when compared to the naive baseline. This result is particularly true of the learned methods, which may also be learning some amount of coreference themselves (as noted by Elson and McKeown (2010)). The rule-based system did not benefit as much, and in some cases performed worse, which was a consequence of the large number of common noun candidates, which often appeared between a quote and its speaker.

With the canonical-based alignment (Table 4) the naive coreference was actually better for quote attribution on the SMHC than the coreference systems, while the P&P results show that Reconcile with the simple binary model outperformed the other combinations. In some respects this is counter to intuition, as the coreference systems are designed for news text and appeared to produce poor results for literature. As noted earlier, the coreference systems tended to over-cluster mentions that shared a family name, even if they had distinct honorifics, which for P&P caused the systems to over-cluster the Bennets, who do most of the talking. This actually causes the quote attribution results to go up, as the alignment methods are imperfect. The naive system does not make the mistake of over-clustering based on family name, and so performs worse with this metric.

The poor results by Stanford and Reconcile on the SMHC are largely caused by their tendency to avoid clustering common nominal mentions with proper nominal mentions. This means that while the correct choice will be a chain containing a proper nominal mention, the quote attribution systems using the candidates from Stanford and Reconcile will have a number of candidate chains that contain only common nominal mentions. As there are no features that allow the quote attribution systems to distinguish these chains from any other chains, they are not able to avoid choosing them. While fixing this problem would be straightforward, it does illustrate that naive use of coreference systems can hurt performance.

8 Coreference Error Analysis

In order to understand some of the problems that were occurring with the coreference systems, we examined some of the main cases of errors. The first problem we identify is that there are a large number of chains with a single mention whose token is POS tagged as a pronoun. Reconcile had the largest number of these with 13,938 (35% of the extracted pronouns) on LIT and 5,501 (33% of the pronouns) on P&P. This is consistent with the result in Kummerfeld and Klein (2013) which finds a large number of missing mentions from Reconcile's output. This problem is particularly acute for quote attribution, as there are a large number of quotes that are directly attributed to a pronoun.

Stanford does better on this problem, having only 1,238 singleton pronouns on LIT and 361 on P&P, of which only 154 and 43, are gendered. Stanford deterministically assigns pronouns to the closest compatible mention in the preceding three sentences and it seems that this is a better way of modelling pronoun discourse. This is in line with Denis and Baldridge (2008)'s claim that the resolution of the different mention types could be more successfully handled with a series of classifiers. However, of these 1,238, 549 are forms of 'you', which suggests that Stanford's discourse sieve needs to be extended to handle the complexities of literature beyond newswire and the conversational data in OntoNotes (Pradhan et al., 2011).

Another major source of errors that we see when manually inspecting the data is conflation of chains corresponding to characters which share a family name, such as the 'Miss Bennet's' and their parents from P&P. To quantify this, we extract all the honorifics within a chain and report cases where a chain is assigned more than one honorific. For Stanford 1.7% of the mentions in LIT and 10.0% of the mentions in P&P are in chains with mixed honorifics, with the majority of the clashes coming from chains including honorifics for both genders. Reconcile makes a similar number of errors with 1.9% of mentions in LIT and 9.9% of mentions in P&P containing clashing honorifics.

9 Conclusions

In this work we addressed the problem of inconsistent candidate speakers within quote attribution corpora. To achieve this we ran three coreference resolution systems over the four corpora, and aligned the gold-standard speakers with chains produced by the coreference systems. This allowed us to more consistently compare the results of three quote attribution methods across the corpora, and additionally provided a more realistic setting for evaluating those methods.

We were also able to use quote attribution as an extrinsic evaluation of coreference resolution. While the speaker alignment methods make it possible to cheat the task, the results are nonetheless informative, and give an indication of how well coreference resolution performs in the literature domain, which has not been assessed with other metrics due to a lack of annotated data.

Future work will include examining the effect of quote extraction on these results, so that the full pipeline effect can be measured. It will also include investigation of features for quote attribution that utilise the information provided by coreference systems. In particular, the number and type of mentions within coreference chains clearly has an impact on the likelihood of them representing a speaker. Lastly, we suggest that coreference systems could be improved by ensuring that honorifics are consistent.

Acknowledgements

O'Keefe was supported by a University of Sydney Merit scholarship and a Capital Markets CRC top-up. This work was supported by ARC Discovery grant DP1097291 and the Capital Markets CRC Computable News project.

References

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

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

David Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pages 1013–1019.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with Wikipedia. *Artificial Intelligence*, 194(0):130–150.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1312–1320. Association for Computational Linguistics.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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 the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.

Ruslan Mitkov, Richard Evans, Constantin Orsan, LeAn Ha, and Viktor Pekar. 2007. Anaphora resolution: To what extent does it help NLP applications? In *Anaphora: Analysis, Algorithms and Applications*, pages 179–190. Springer.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Tim O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799. Association for Computational Linguistics.

Silvia Pareti. 2011. Annotating attribution relations and their features. In *Proceedings of the fourth workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 19–20. ACM.

Silvia Pareti. 2012. A database of attribution relations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3213–3217.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2006. Annotating attribution in the Penn Discourse TreeBank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 31–38.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning.

2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977. Association for Computational Linguistics.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with reconcile. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 156–161. Association for Computational Linguistics.

Veselin Stoyanov and Jason Eisner. 2012. Easy-first coreference resolution. In *Proceedings of the 24th Internation Conference on Computational Linguistics 2012*.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52. Association for Computational Linguistics.

Ralph Weischedel and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus. *Linguistic Data Consortium, Philadelphia*.