Automatically Detecting and Attributing Indirect Quotations

Silvia Pareti[◇*] **Tim O'Keefe**[†*] **Ioannis Konstas**[◇] **James R. Curran**[†] **Irena Koprinska**[†]

[◇]ILCC, School of Informatics
University of Edinburgh
United Kingdom
{s.pareti,i.konstas}@sms.ed.ac.uk

[†]ə-lab, School of IT
University of Sydney
NSW 2006, Australia
{tokeefe,james,irena}@it.usyd.edu.au

Abstract

Direct quotations are used for opinion mining and information extraction as they have an easy to extract span and they can be attributed to a speaker with high accuracy. However, simply focusing on direct quotations ignores around half of all reported speech, which is in the form of indirect or mixed speech. This work presents the first large-scale experiments in indirect and mixed quotation extraction and attribution. We propose two methods of extracting all quote types from news articles and evaluate them on two large annotated corpora, one of which is a contribution of this work. We further show that direct quotation attribution methods can be successfully applied to indirect and mixed quotation attribution.

1 Introduction

Quotations are crucial carriers of information, particularly in news texts, with up to 90% of sentences in some articles being reported speech (Bergler et al., 2004). Reported speech is a carrier of evidence and factuality (Bergler, 1992; Saurí and Pustejovsky, 2009), and as such, text mining applications use quotations to summarise, organise and validate information. Extraction of quotations is also relevant to researchers interested in media monitoring.

Most quotation attribution studies (Pouliquen et al., 2007; Glass and Bangay, 2007; Elson and McKeown, 2010) thus far have limited their scope to direct quotations (Ex.1a), as they are delimited by quotation marks, which makes them easy to extract. However, annotated resources suggest that direct quotations represent only a limited portion of all quotations, i.e., around 30% in the Penn Attribution Relation Corpus (PARC), which covers Wall Street Journal articles, and 52% in the Sydney Morning Herald Corpus (SMHC), with the remainder being indirect (Ex.1c) or mixed (Ex.1b) quotations. Retrieving only direct quotations can miss key content that can change the interpretation of the quotation (Ex. 1b) and will entirely miss indirect quotations.

(1) a. *"For 10 million, you can move $100 million of stocks,"* a specialist on the Big Board gripes. *"That gives futures traders a lot more power."*
b. *Police would only apply for the restrictions when "we have a lot of evidence that late-night noise... is disturbing the residents of that neighbourhood"*, Superintendent Tony Cooke said.
c. Mr Walsh said *Rio was continuing to hold discussions with its customers to arrive at a mutually agreed price*.

Previous work on extracting indirect and mixed quotations has suffered from a lack of large-scale data, and has instead used hand-crafted lexica of reporting verbs with rule-based approaches. The lack of data has also made comparing the relative merit of these approaches difficult, as existing evaluations are small-scale and do not compare multiple methods on the same data.

In this work we address this lack of clear, comparable results by evaluating two baseline meth-

*These authors contributed equally to this work.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 989–999,
Seattle, Washington, USA, 18-21 October 2013.

	Method	Language	Test Size (quotations)	Results P	R
Krestel et al. (2008)	hand-built grammar	English	133	74%	99%
Sarmento and Nunes (2009)	patterns over text	Portuguese	570	88%	5%[1]
Fernandes et al. (2011)	ML and regex	Portuguese	205	64%[2]	67%[2]
de La Clergerie et al. (2011)	patterns over parse	French	40	87%	70%
Schneider et al. (2010)	hand-built grammar	English	N/D	56%[2]	52%[2]

Table 1: Related work on direct, indirect and mixed quotation extraction. Note that they are not directly comparable as they apply to different languages and greatly differ in evaluation style and size of test set. [1] Figure estimated by the authors for extracting 570 quotations from 26k articles. [2] Results are for quotation extraction and attribution jointly.

ods against both a token-based approach that uses a Conditional Random Field (CRF) to predict IOB labels, and a maximum entropy classifier that predicts whether parse nodes are quotations or not. We evaluate these approaches on two large-scale corpora from the news domain that together include over 18,000 quotations. One of these corpora (SMHC) is a a contribution of this work, while our results are the first presented on the other corpus (PARC). Instead of relying on a lexicon of reporting verbs, we develop a classifier to detect verbs introducing a quotation. To inform future research we present results for direct, indirect, and mixed quotations, as well as overall results.

Finally, we use the direct quotation attribution methods described in O'Keefe et al. (2012) and show that they can be successfully applied to indirect and mixed quotations, albeit with lower accuracy. This leads us to conclude that attributing indirect and mixed quotations to speakers is harder than attributing direct quotations.

With this work, we set a new state of the art in quotation extraction. We expect that the main contribution of this work will be that future methods can be evaluated in a comparable way, so that the relative merit of various approaches can be determined.

2 Background

Pareti (2012) defines an attribution as having a *source* span, a *cue* span, and a *content* span:

Source is the span of text that indicates who the content is attributed to, e.g. 'president Obama', 'analysts', 'China', 'she'.

Cue is the lexical anchor of the attribution relation, usually a verb, e.g. 'say', 'add', 'quip'.

Content is the span of text that is attributed.

Based on the type of attitude the source expresses towards a proposition or eventuality, attributions are subcategorised (Prasad et al., 2006) into *assertions* (Ex.2a) and *beliefs* (Ex.2b), which imply different degrees of commitment, *facts* (Ex.2c), expressing evaluation or knowledge, and *eventualities* (Ex.2d), expressing intention or attitude.

(2) a. Mr Abbott said *that he will win the election.*
b. Mr Abbott thinks *he will win the election.*
c. Mr Abbott knew *that Gillard was in Sydney.*
d. Mr Abbott agreed *to the public sector cuts.*

Only assertion attributions necessarily imply a speech act. Their *content* corresponds to a quotation span and their *source* is generally referred to in the literature as the *speaker*. Direct, indirect and mixed quotations differ in the degree of factuality they entail, since the former are by convention interpreted as a verbatim transcription of an utterance whereas indirect and the non-quoted portion of mixed quotations can be paraphrased forms of the original wording, and are thus filtered by the writer's perspective.

The first speaker attribution systems (Zhang et al., 2003; Mamede and Chaleira, 2004; Glass and Bangay, 2007) originate from the narrative domain and were concerned with the identification of different characters for speech synthesis applications. Direct quotation attribution, with direct quotations being given or extracted heuristically, has been the focus of further studies in both the narrative (Elson and McKeown, 2010) and news (Pouliquen et al., 2007; Liang et al., 2010) domains. The few studies that

have addressed the extraction and attribution of indirect and mixed quotations are discussed below.

Krestel et al. (2008) developed a quotation extraction and attribution system that combines a lexicon of 53 common reporting verbs and a hand-built grammar to detect constructions that match 6 general lexical patterns. They evaluate their work on 7 articles from the Wall Street Journal, which contain 133 quotations, achieving macro-averaged Precision (P) of 99% and Recall (R) of 74% for quotation span detection. PICTOR (Schneider et al., 2010) relies instead on a context-free grammar for the extraction and attribution of quotations. PICTOR yielded 75% P and 86% R in terms of words correctly ascribed to a quotation or speaker, while it achieved 56% P and 52% R when measured in terms of completely correct quotation-speaker pairs.

SAPIENS (de La Clergerie et al., 2011) extracts quotations from French news, by using a lexicon of reporting verbs and syntactic patterns to extract the complement of a reporting verb as the quotation span and its subject as the source. They evaluated 40 randomly sampled quotations and found that their system made 32 predictions and correctly identified the span in 28 of the 40 cases. Verbatim (Sarmento and Nunes, 2009) extracts quotations from Portuguese news feeds by first finding one of 35 speech verbs and then matching the sentence to one of 19 patterns. Their manual evaluation shows that 11.9% of the quotations Verbatim finds are errors and that the system identifies approximately one distinct quotation for every 46 news articles.

The system presented by Fernandes et al. (2011) also works over Portuguese news. Their work is the closest to ours as they partially apply supervised machine learning to quotation extraction. Their work introduces GloboQuotes, a corpus of 685 news items containing 1,007 quotations of which 802 were used to train an Entropy Guided Transformation Learning (ETL) algorithm (dos Santos and Milidiú, 2009). They treat quotation extraction as an IOB labelling task, where they use ETL with POS and NE features to identify the beginning of a quotation, while the inside and outside labels are found using regular expressions. Finally they use ETL to attribute quotations to their source. The overall system achieves 64% P and 67% R.

We have summarised these approaches in Table 1, which shows that the majority of evaluations thus far have been small-scale. Furthermore, the published results do not include any comparisons with previous work, which prevents a quantitative comparison of the approaches, and they do not include results broken down by whether the quotation is direct, indirect, or mixed. It is these issues that motivate our work.

	SMHC		PARC	
	Corpus	Doc	Corpus	Doc
Docs	965	-	2,280	-
Tokens	601k	623.3	1,139k	499.9
Quotations	7,991	8.3	10,526	4.6
Direct	4,204	4.4	3,262	1.4
Indirect	2,930	3.0	5,715	2.5
Mixed	857	0.9	1,549	0.6

Table 2: Comparison of the SMHC and PARC corpora, reporting their document and token size and per-type occurrence of quotations overall and per document (average).

3 Corpora

We perform our experiments over two large corpora from the news domain.

3.1 Penn Attribution Relations Corpus (PARC)

Our first corpus (Pareti, 2012), which we will refer to as PARC, is a semi-automatically built extension to the attribution annotations included in the PDTB (Prasad et al., 2008). The corpus covers 2,280 Wall Street Journal articles and contains annotations of assertions, beliefs, facts, and eventualities, which are altogether referred to as attribution relations (ARs). For this work we use only the assertions, as they correspond to quotations (direct, indirect and mixed). The drawback of this corpus is that it is not yet fully annotated, i.e., it comprises positive and unlabelled data.

The corpus includes a test set of 14 articles that are fully annotated, which enables us to properly evaluate our work and estimate that a proportion of 30-50% of ARs are unlabelled in the rest of the corpus. The test set was manually annotated by two expert annotators. The annotators identified 491 ARs, of which 22% were nested within another AR, with

an agreement score of 87%[1]. The agreement for the selection of the content and source spans of commonly annotated ARs was 95% and 94% respectively. In this work we address only non-embedded assertions, so the final test-set includes 267 quotes, totalling 321 non-discontinuous gold spans.

3.2 Sydney Morning Herald Corpus (SMHC)

We based our second corpus on the existing annotations of direct quotations within Sydney Morning Herald articles presented in O'Keefe et al. (2012). In that work we defined direct quotations as any text between quotation marks, which included the directly-quoted portion of mixed quotations, as well as scare quotes. Under that definition direct quotations could be automatically extracted with very high accuracy, so annotations in that work were over the automatically extracted direct quotations. As part of this work one annotator removed scare quotes, updated mixed quotations to include both the directly and indirectly quoted portions, and added whole new indirect quotations. The annotation scheme was developed to be comparable to the scheme used in the PARC corpus (Pareti, 2012), although the SMHC corpus only includes assertions and does not annotate the lexical *cue*.

The resulting corpus contains 7,991 quotations taken from 965 articles from the 2009 Sydney Morning Herald (we refer to this corpus as SMHC). The annotations in this corpus also include the speakers of the quotations, as well as gold standard Named Entities (NEs). We use 60% of this corpus as training data (4,872 quotations), 10% as development data (759 quotations), and 30% as test data (2,360 quotations). Early experiments were conducted over the development data, while the final results were trained on both the training and development sets and were tested on the unseen test data.

3.3 Comparison

Table 2 shows a comparison of the two corpora and the quotations annotated within them. SMHC has a higher density of quotations per document, 8.3 vs. 4.6 in PARC, since articles are fully annotated and were selected to contain at least one quotation. PARC is instead only partially annotated and comprises articles with no quotations. Excluding null-quotation articles from PARC, the average incidence of annotated quotations per article raises to 7.1. The corpora also differ in quotation type distribution, with direct quotations being largely predominant in SMHC while indirect are more common in PARC.

[1]The agreement was calculated using the *agr* metric described in Wiebe and Riloff (2005) as the proportion of commonly annotated ARs with respect to the ARs identified overall by Annotator A and Annotator B respectively

	P	R	F
B_{say}	94.4	43.5	59.5
B_{list}	75.4	71.1	73.2
k-NN	88.9	72.6	79.9

Table 3: Results for the k-NN verb-cue classifier. B_{say} classifies as verb-cue all instances of say while B_{list} marks as verb-cues all verbs from a pre-compiled list in Krestel et al. (2008).

4 Experimental Setup

4.1 Quotation Extraction

Quotation extraction is the task of extracting the *content span* of all of the direct, indirect, and mixed quotations within a given document. More precisely, we consider quotations to be acts of communication, which correspond to *assertions* in Pareti (2012). Some quotations have content spans that are split into separate, non-adjacent spans, as in example (1a). Ideally the latter span should be marked as a continuation of a quotation, however we consider this to be out of scope for this work, so we treat each span as a separate quotation.

4.2 Preprocessing

As a pre-processing step, both corpora were tokenised and POS tagged, and the potential speakers anonymised to prevent over-fitting. We used the Stanford factored parser (Klein and Manning, 2002) to retrieve both the Stanford dependencies and the phrase structure parse. Quotation marks were normalised to a single character, as the quotation direction is often incorrect for multi-paragraph quotations.

4.3 Verb-cue Classifier

Verbs are by far the most common introducer of a quotation. In PARC verbs account for 96% of all

cues, the prepositional phrase *according to* for 3%, with the remaining 1% being nouns, adverbials and prepositional groups. Attributional verbs are not a closed set, they can vary across styles and genres, and their attributional use is highly dependent on the context in which they occur. It is therefore not possible to simply rely on a pre-compiled list of common speech verbs. Quotations in PARC are introduced by 232 verb types, 87 of which are unique occurrences. Not all of the verbs are speech verbs, for example *add*, which is the second most frequent after *say*, or the manner verb *gripe* (Ex.1a).

We used the attributional cues in the PARC corpus to develop a separate component of our system to identify attribution verb-cues. The classifier predicts whether the head of each verb group is a verb-cue using the k-nearest neighbour (k-NN) algorithm, with k equal to 3. The classifier uses 20 feature types, including:

- Lexical (e.g. token, lemma, adjacent tokens)
- VerbNet classes membership
- Syntactic (e.g. node-depth in the sentence, parent and sibling nodes)
- Sentence features (e.g. distance from sentence start/end, within quotation markers).

We compared the system to one baseline, B_{say}, that marks every instance of *say* as a verb-cue, and another, B_{list}, that marks every instance of a verb that is on the list of 53 verbs presented in Krestel et al. (2008). We tested the system on the test set for PARC, which contains 1809 potential verb-cues, of which 354 are positive and 1455 are negative.

The results in Table 3 show that the verb-cue classifier can outperform expert-derived knowledge. The classifier was able to identify verb-cues with P of 88.9% and R of 72.6%. While frequently occurring verbs are highly predictive, the inclusion of VerbNet classes (Schuler, 2005) and contextual features allows for a more accurate classification of polysemous and unseen verbs.

Since PARC contains labelled and unlabelled attributions, which is detrimental for training, we used the verb-cue classifier to identify in the corpus sentences that we suspected contained an unlabelled attribution. Sentences containing a verb classified as a cue that do not contain a quotation were removed from the training set for the quotation extraction model.

4.4 Evaluation

We use two metrics, listed below, for evaluating the quotation spans predicted by our model against the gold spans from the annotation.

Strict The first is a strict metric where a predicted span is only considered to be correct if it exactly matches a span from the gold standard. The standard precision, recall, and F-score can be calculated using this definition of correctness. The drawback of this strict score is that if a prediction is incorrect by as little as one token it will be considered completely incorrect.

Partial We also consider an overlap metric (Hollingsworth and Teufel, 2005), which allows partially correct predictions to be proportionally counted. Precision (P), recall (R), and F-score for this method are:

$$P = \frac{\sum_{g \in gold} \sum_{p \in pred} overlap(g, p)}{|pred|} \quad (1)$$

$$R = \frac{\sum_{g \in gold} \sum_{p \in pred} overlap(p, g)}{|gold|} \quad (2)$$

$$F = \frac{2PR}{(P + R)} \quad (3)$$

Where $overlap(x, y)$ returns the proportion of tokens of y that are overlapped by x. For each of these metrics we report the micro-average, as the number of quotations in each document varies significantly. When reporting P for the typewise results we restrict the set of predicted quotations to only those with the requisite type, while still considering the full set of gold quotations. Similarly, when calculating R we restrict the set of gold quotations to only those with the required type.

4.5 Baselines

We have developed two baselines inspired by the current lexical/syntactic pattern-based approaches in the literature, which combine speech verbs and hand-crafted rules.

B_{lex} Lexical: cue verb + the longest of the spans before or after it until the sentence boundary.

B_{syn} Syntactic: cue verb + verb syntactic object. B_{syn} is close to the model in de La Clergerie et al. (2011).

Instead of relying on a lexicon of verbs, our baselines use those identified by the verb-cue classifier. As direct quotations are not always explicitly introduced by a cue-verb, we defined a separate baseline with a rule-based approach (B_{rule}) that returns text between quotation marks that has at least 3 tokens, and where the non-stopword and non-proper noun tokens are not all title cased. In our full results we apply each method along with B_{rule} and greedily take the longest predicted spans that do not overlap.

5 Supervised Approaches

We present two supervised approaches to quotation extraction, which operate over the tokens and the phrase-structure parse nodes respectively. Despite the difference in the item being classified, these approaches have some common features:

Lexical: unigram and bigram versions of the token, lemma, and POS tags within a window of 5 tokens either side of the target, all indexed by position.

Sentence: features indicating whether the sentence contains a quotation mark, a NE, a verb-cue, a pronoun, or any combination of these. There is also a sentence length feature.

Dependency: relation with parent, relations with any dependants, as well as versions of these that include the head and dependent tokens.

External knowledge: position-indexed features for whether any of the tokens in the sentence match a known role, organisation, or title. The titles come from a small hand-built list, while the role and organisation lists were built by recursively following the WordNet (Fellbaum, 1998) hyponyms of person and organization respectively.

Other: features for whether the target is within quotation marks, and whether there is a verb-cue near the end of the sentence.

		Strict			Partial		
		P	R	F	P	R	F
PARC	B_{rule}	75	94	83	96	94	95
	Token	97	91	94	98	97	97
SMHC	B_{rule}	87	93	90	98	94	96
	Token	94	90	92	99	97	98

Table 4: PARC and SMHC results on direct quotations. The token based approach is trained and tested on all quotations.

5.1 Token-based Approach

The token-based approach treats quotation extraction as analogous to NE tagging, where there are a sequence of tokens that need to be individually labelled. Each token is given either an I, an O, or a B label, where B denotes the first token in a quotation, I denotes the token is inside a quotation, and O indicates that the token is not part of a quotation. For NE tagging it is common to use a sentence as a single sequence, as NEs do not cross sentence boundaries. This does not work for quotations, as they can cross sentence and even paragraph boundaries. As such, we treat the entire document as a single sequence, which allows the predicted quotations to span both sentence and paragraph bounds.

We use a linear chain Conditional Random Field (CRF)[2] as the learning algorithm, with the common features listed above, as well as the following features:

Verb: features indicating whether the current token is a (possibly indirect) dependent of a verb-cue, and another for whether the token is at the start of a constituent that is a dependent of a verb-cue.

Ancestor: the labels of all constituents that contain the current token in their span, indexed by their depth in the parse tree.

Syntactic: the label, depth, and token span size of the highest constituent where the current token is the left-most token in the constituent, as well as its parent, and whether either of those contains a verb-cue.

[2]http://www.chokkan.org/software/crfsuite/

	Indirect			Mixed			All[1]		
Strict	P	R	F	P	R	F	P	R	F
B_{lex}	34	32	33	17	26	20	46	44	45
B_{syn}	78	46	58	61	40	49	80	63	70
Token	66	54	59	55	58	56	76	70	73
Constituent	61	50	55	50	38	43	70	64	67
Constituent$_G$	66	42	51	68	49	57	76	62	68
Partial	P	R	F	P	R	F	P	R	F
B_{lex}	56	66	61	78	79	78	73	79	76
B_{syn}	89	58	70	88	75	81	92	74	82
Token	79	74	76	85	90	87	87	86	87
Constituent	78	67	72	84	82	83	86	80	83
Constituent$_G$	80	54	65	90	80	85	90	74	81

Table 5: Results on PARC. [1]*All* reports the results over all quotations (direct, indirect and mixed). For the baselines, this is a combination of the strategy in B_{lex} or B_{syn} with the rules for direct quotations. Constituent$_G$ shows the results for the constituent model using the gold parse.

5.2 Constituent-based Approach

The constituent approach classifies whole phrase structure nodes as either *quotation* or *not a quotation*. Ideally each quotation would match exactly one constituent, however this is not always the case in our data. In cases without an exact match we label every constituent that is a subspan of the quotation as a *quotation* as long as it has a parent that is not a subspan of the quotation. In these cases multiple nodes will be labelled *quotation*, so a post-processing step is introduced that rebuilds quotations by merging predicted spans that are adjacent or overlapping within a sentence. Restricting the merging process this way loses the ability to predict quotations that cover more than a sentence, but without this restriction too many predicted quotations are erroneously merged.

This approach uses a maximum entropy classifier[3] with $L1$ regularisation. In early experiments we found that the constituent-based approach performed poorly when trained on all quotations, so for these experiments the constituent classifier is trained only on indirect and mixed quotations. The classifier uses the common features listed above as well as the following features:

Span: length of the span, features for whether there is a verb or a NE.

Node: the label, number of descendants, number of ancestors, and number of children of the target.

Context: dependency, node, and span features for the parent and siblings of the target.

In addition the lexical features described earlier are applied to both the start and end tokens of the node's span, as well as the highest token in the dependency parse that is within the span.

6 Results

6.1 Direct Quotations

Table 4 shows the results for predicting direct quotations on PARC and SMHC. In both corpora and with both metrics the token-based approach outperforms B_{rule}. Although direct quotations should be trivial to extract, and a simple system that returns the content between quotation marks should be hard to beat, there are two main factors that confound the rule-based system.

The first is the presence of mixed quotations, which is most clearly demonstrated in the difference between the strict precision scores and the partial precision scores for B_{rule}. B_{rule} will find all of the directly-quoted portions of mixed quotes, which do not exactly match a quotation, and so will receive a low precision score with the strict metric. However the partial overlap score will reward these

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

	Indirect			Mixed			All[1]		
Strict	P	R	F	P	R	F	P	R	F
B_{lex}	37	42	40	15	36	21	50	50	50
B_{syn}	63	49	55	67	36	47	82	72	76
Token	69	53	60	80	91	85	82	75	78
Constituent	54	49	51	64	42	51	77	72	75
Partial	P	R	F	P	R	F	P	R	F
B_{lex}	52	68	59	87	77	82	77	84	81
B_{syn}	75	59	66	89	66	76	91	80	85
Token	82	67	74	88	84	86	92	86	89
Constituent	77	63	69	91	75	82	91	82	86

Table 6: Results on SMHC. [1]*All* reports the results over all quotations (direct, indirect and mixed). For the baselines, this is a combination of the strategy in B_{lex} or B_{syn} with the rules for direct quotations.

predictions, as they do partially match a quote, so there is a large difference in those scores. Note that the reduced strict score does not occur for the token method, which correctly identifies mixed quotations.

The other main issue is the presence of quotation marks around items such as book titles and scare quotes (i.e. text that is in quotation marks to distance the author from a particular wording or claim). In Section 4.5 we described the methods that we use to avoid scare quotes and titles, which are rule-based and imperfect. While these methods increase the overall F-score of B_{rule}, they do have a negative impact on recall, which is why the recall is lower than might be expected. These results demonstrate that although direct quotations can be accurately extracted with rules, the accuracy will be lower than might be anticipated and the returned spans will include a number of mixed quotations, which will be missing some content.

6.2 Indirect and Mixed Quotations

The token approach was also the most effective method for extracting indirect and mixed quotations as Tables 5 and 6 show. Indirect quotations were extracted with strict F-scores of 59% and 60% and partial F-scores of 76% and 74% in PARC and SMHC respectively, while mixed quotes were found with strict F-scores of 56% and 85% and partial F-scores of 87% and 86%.

Although there is a strong interconnection between syntax and attribution, results for B_{syn} show that merely considering attribution as a syntactic relation (Skadhauge and Hardt, 2005) has a large impact on recall: only a subset of inter-sentential quotations can be effectively matched by verb complement boundaries.

The constituent model yielded lower results than the token one, and in particular it greatly lowered the recall of mixed quotations in both corpora. Since the model heavily relies on syntax, it is particularly affected by errors made by the parser. The conjunction *and* in Example 3 is incorrectly attached by the parser to the cue *said*, leading the classifier to identify two separate spans. In order to verify the impact of incorrect parsing on the model, we ran the constituent model using gold standard parses for PARC. This resulted in an increase in strict P and increased the F-score for mixed quotations to 57%, similarly to the score achieved by the token model. However, it surprisingly negatively affected R for indirect quotations.

(3) Graeme Hugo, said *strong links between Australia's 700,000 ethnic Chinese and China could benefit both countries* ***and*** *were unlikely to pose a threat*.

The tables also report results for the extraction of all quotations, irrespective of their type. For this score, the baseline models for indirect and mixed quotations are combined with B_{rule} for direct quotations.

6.3 Model Comparison

We designed the features for the token and constituent models to be largely similar. This al-

lows us to conclude that the difference in performance between the token and constituent models is largely driven by the class labelling and learning method. Overall, the token-based approach outperformed both the baselines and the constituent method. Qualitatively we found that the token-based approach was making reasonable predictions most of the time, but would often fail when a quotation was attributed to a speaker through a parenthetical clause, as in Example 4.

(4) *Finding lunar ice*, said Tidbinbilla's spokesman, Glen Nagle, *would give a major boost to NASA's hopes of returning humans to the moon by 2020.*

The token-based approach has a reasonable balance of the various label types, and benefits from a decoding step that allows it to make trade-offs between good local decisions and a good overall solution. By comparison, the constituent-based approach has a large class imbalance, as there are many more negative (i.e. not quotation) parse nodes than there are positive, which makes finding a good decision boundary difficult. We experimented with reducing the number of negative nodes to consider, but found that the overall F-score was equivalent or worse, largely driven by a drop in recall. We also found that in many cases the constituent-approach predicted quotes that were too short, or that were only the second half of a conjunction, without the first half being labelled. We expect that these issues would be corrected with the addition of a decoding step, that forces the classifier to make a good global decision.

7 Speaker Attribution

While the focus of this paper is on extracting quotations, we also present results on finding the speaker of each quotation. As discussed in Section 2, quotation attribution has been addressed in the literature before, including some work that includes large-scale data (Elson and McKeown, 2010). However, the large-scale evaluations that exist cover only direct quotations, whereas we present results for direct, indirect, and mixed quotations.

For this evaluation we use four of the methods that were introduced in O'Keefe et al. (2012). The first is a simple rule-based approach (Rule) that returns the entity closest to the speech verb nearest the quotation, or if there is no such speech verb then the entity nearest the end of the quotation. The second method uses a CRF which is able to choose between up to 15 entities that are in the paragraph containing the quotation or any preceding it. The third method (No seq.) is a binary MaxEnt classifier that predicts whether each entity is the speaker or not the speaker, with the entity achieving the highest speaker probability predicted. In O'Keefe et al. (2012) this model achieved the best results on the direct quotations in SMHC, despite not using the sequence features or decoding methods that were available to other models. The final method that we evaluate (Gold) is the approach that uses sequence features that use the gold-standard labels from previous decisions. As noted by O'Keefe et al., this method is not realisable in practise, however we include these results so that we can reassess the claims of O'Keefe et al. when direct, indirect, and mixed quotations are included. For our results to be comparable we use the list of speech verbs that was presented in Elson and McKeown (2010) and used in O'Keefe et al. (2012).

Table 7 shows the accuracy of the two methods on both PARC and SMHC, broken down by the type of the quotation. The first observation that we make about these results in comparison to the O'Keefe et al. results, is that the accuracy is generally lower, even for direct quotations. This discrepancy is caused by differences in our data compared to theirs, notably that the sequence of quotations is altered in ours by the introduction of indirect quotations, and that some of the direct quotations that they evaluated would be considered mixed quotations in our corpora. The rule based method performs particularly poorly on PARC, which is likely caused by the relative scarcity of direct quotations and the fact that it was designed for direct quotations only. Direct quotations are much more frequent in SMHC, so the rules that rely on the sequence of speakers would likely perform relatively better than on PARC.

While the approach using gold-standard sequence features unsurprisingly performed the best, the most straightforward learned model (No seq.), trained without any sequence information, equalled or outperformed the two other non-gold approaches for all quotation types on both corpora. This indicates that the CRF model evaluated here was not able to effec-

Corpus	Method	Dir.	Ind.	Mix.	All
PARC	Rule	70	60	47	62
	CRF	82	68	65	73
	No seq.	85	74	65	77
	Gold	88	79	74	82
SMHC	Rule	89	76	78	84
	CRF	83	72	71	78
	No seq.	91	79	81	87
	Gold	93	81	83	89

Table 7: Speaker attribution accuracy results for both corpora over gold standard quotations.

tively use the sequence information that is present.

8 Conclusion

In this work we have presented the first large-scale experiments on the entire quotation extraction and attribution task: evaluating the extraction and attribution of direct, indirect and mixed quotations over two large news corpora. One of these corpora (SMHC) is a novel contribution of this work, while our results are the first presented for the other corpus (PARC). This work has shown that while rule-based approaches that return the object of a speech verb are indeed effective, they are outperformed by supervised systems that can take advantage of additional evidence. We also show that state-of-the-art quotation attribution methods are less accurate on indirect and mixed quotations than they are on direct quotations.

Future work will include extending these methods to extract all attributions, i.e. beliefs, eventualities, and facts, as well as the source spans. We will also evaluate the effect of adding a decoding step to the constituent approach. This work provides an accurate and complete quotation extraction and attribution system that can be used for a wide range of tasks in information extraction and opinion mining.

Acknowledgements

We would like to thank Bonnie Webber for her feedback and assistance. Pareti has been supported by a Scottish Informatics & Computer Science Alliance (SICSA) studentship; O'Keefe has been supported by a University of Sydney Merit scholarship and a Capital Markets CRC top-up scholarship. This work has been supported by ARC Discovery grant DP1097291 and the Capital Markets CRC Computable News project.

References

Sabine Bergler. 1992. *Evidential analysis of reported speech*. Ph.D. thesis, Brandeis University.

Sabine Bergler, Monia Doandes, Christine Gerard, and René Witte. 2004. Attributions. In *Exploring Attitude and Affect in Text: Theories and Applications*, Technical Report SS-04-07, pages 16–19. Papers from the 2004 AAAI Spring Symposium.

Eric de La Clergerie, Benoit Sagot, Rosa Stern, Pascal Denis, Gaelle Recource, and Victor Mignot. 2011. Extracting and visualizing quotations from news wires. *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 522–532.

Cícero Nogueira dos Santos and Ruy Luiz Milidiú. 2009. Entropy guided transformation learning. In *Foundations of Computational, Intelligence Volume 1*, Studies in Computational Intelligence, pages 159–184. Springer.

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

Christine Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press Cambridge, MA.

William Paulo Ducca Fernandes, Eduardo Motta, and Ruy Luiz Milidiú. 2011. Quotation extraction for portuguese. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011)*, pages 204–208.

Kevin Glass and Shaun Bangay. 2007. A naive salience-based method for speaker identification in fiction books. In *Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA07)*, pages 1–6.

Bill Hollingsworth and Simone Teufel. 2005. Human annotation of lexical chains: Coverage and agreement measures. In *ELECTRA Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (Beyond Bag of Words)*, page 26.

Dan Klein and Christopher D Manning. 2002. Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems*, pages 3–10.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

Jisheng Liang, Navdeep Dhillon, and Krzysztof Koperski. 2010. A large-scale system for annotating and querying quotations in news feeds. In *Proceedings of the 3rd International Semantic Search Workshop*, pages 1–5.

Nuno Mamede and Pedro Chaleira. 2004. Character identification in children stories. *Advances in Natural Language Processing*, pages 82–90.

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.

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

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492.

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.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0 annotation manual. In *Technical report, University of Pennsylvania: Institute for Research in Cognitive Science*.

Luis Sarmento and Sergio Nunes. 2009. Automatic extraction of quotes and topics from news feeds. In *4th Doctoral Symposium on Informatics Engineering*.

Roser Saurí and James Pustejovsky. 2009. Factbank: A corpus annotated with event factuality. *In Language Resources and Evaluation*, pages 227–268.

Nathan Schneider, Rebecca Hwa, Philip Gianfortoni, Dipanjan Das, Michael Heilman, Alan W. Black, Frederik L. Crabbe, and Noah A. Smith. 2010. Visualizing topical quotations over time to understand news discourse. Technical report, Carnegie Mellon University.

Karin K. Schuler. 2005. *Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, Faculties of Computer and Information Science of the University of Pennsylvania.

Peter R. Skadhauge and Daniel Hardt. 2005. Syntactic identification of attribution in the RST treebank. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora*.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing*, pages 486–497. Springer.

Jason Zhang, Alan Black, and Richard Sproat. 2003. Identifying speakers in children's stories for speech synthesis. In *Proceedings of EUROSPEECH*, pages 2041–2044.