PARC 3.0: A Corpus of Attribution Relations

Silvia Pareti[+]
School of Informatics
The University of Edinburgh, UK
S.Pareti@sms.ed.ac.uk

Abstract

Quotation and opinion extraction, discourse and factuality have all partly addressed the annotation and identification of Attribution Relations. However, disjoint efforts have provided a partial and partly inaccurate picture of attribution and generated small or incomplete resources, thus limiting the applicability of machine learning approaches. This paper presents PARC 3.0, a large corpus fully annotated with attribution relations (ARs). The annotation scheme was tested with an inter-annotator agreement study showing satisfactory results for the identification of ARs and high agreement on the selection of the text spans corresponding to its constitutive elements: source, cue and content. The corpus, which comprises around 20k ARs, was used to investigate the range of structures that can express attribution. The results show a complex and varied relation of which the literature has addressed only a portion. PARC 3.0 is available for research use and can be used in a range of different studies to analyse attribution and validate assumptions as well as to develop supervised attribution extraction models.

Keywords: attribution, corpora, quotations

1. Introduction

With a vast amount of data being available, in particular through the world wide web, more than ever before users have the chance to access an enormous amount of information. While information per se is a resource, this information overload can hinder our ability to process it and use it to understand issues or make decisions. To manage the vast amount of information available today requires ways to organise, filter and select it. It therefore becomes important to recognise different point of views (e.g. to make a medical or financial decision), monitor the statements of a specific person (e.g. a politician) and identify truthful and reliable information. These tasks require the identification of attribution relations, thus allowing to link the attributed material to the entity representing its source.

Attribution allows to identify what has been attributed to a specific source but also affects how the text itself is perceived. Changes in the source or attributional verb (e.g. *say* vs *suspect* vs *joke*) can affect our perception of the quoted statement.

While research and commercial systems for the automatic identification and extraction of attribution relations have multiplied in recent years (e.g. NewsExplorer[1] (Pouliquen et al., 2007)), several issues are still to be addressed. The applications of such systems are severely limited by low precision and low recall.

The reason for this relatively poor performance is partly to be found in the limited scope of such approaches. Studies on attribution focused either on its overlap and interaction with other linguistic phenomena, such as discourse relation (Carlson and Marcu, 2001; Wolf and Gibson, 2005) and factuality (Saurí and Pustejovsky, 2009), or on specific types of attributions such as direct quotations (Elson and McKeown, 2010) or opinions (Wiebe, 2002). While these studies show that attribution is relevant for different linguist fields, their approaches address only a subset of attribution or rely on small and partially annotated resources. These are inadequate to guide a comprehensive understanding of attribution and drive the development of extraction systems. Since the lack of annotated data hindered the development of supervised computational models, the systems developed had to rely mostly on hand-crafted rules and results could be tested on a small number of examples.

This paper presents PARC 3.0, a large corpus of ARs developed with the intent of supporting a wide range of studies and the training of computational models. The corpus was developed starting from the annotation schema proposed in Pareti and Prodanof (2010) and the partial annotation of attribution included in the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008). The PDTB is a collection of over 2,000 news articles from the Wall Street Journal (WSJ) annotated with discourse connectives and their arguments. Attribution can be found in many types of human communication, whether verbal or not. Different attributive structures and means are used in different genres, including graphical and acoustic clues. The focus of this work is on news not only because of the ubiquity of this phenomenon in the news genre, but also because one of the goals of collecting such resource is to enable studying the effect of attribution on information.

2. Background

While several resources comprise some annotation of attribution, these resources are mostly too small or narrow-scope to be employed to train supervised extraction systems. A comparison of the most relevant corpora is presented in Table 1. Most resources consist of English news texts and corpora can be as small as 40 texts and contain only few attributions. Apart from quotation corpora, attribution is usually not directly annotated, but included as a discourse relation or a carrier of opinion.

The only large resources that comprise attribution annotations are: the PDTB; the Multi-Perspective Question An-

[+]Current affiliation: Google Inc., spareti@google.com
[1]http://emm.newsexplorer.eu/

Corpus	Annotations	Texts	Genre	Language	Type
PDTB (Prasad et al., 2008)	10k	2,159	news	EN	discourse, ARs
RST (Carlson and Marcu, 2001)	small	385	news	EN	discourse
GraphBank (Wolf and Gibson, 2005)	small	135	news	EN	discourse
MPQA (Wiebe, 2002)	-	692	news	EN	opinions
NTCIR (Seki et al., 2008; Seki et al., 2010)	4.5k-9.5k	150-800	news	EN/JA/ZH	opinions
TimeBank (Pustejovsky et al., 2003)	small	183	news	EN	events
CQSA (Elson and McKeown, 2010)	3.5k	11 books	narrative	EN	quotes
ItAC (Pareti, 2009; Pareti and Prodanof, 2010)	461	50	news	IT	ARs
DENews (Li et al., 2012)	315	108	news	DE	opinions
CorpusTCC (Pardo and Nunes, 2003)	185	100	scientific	PT	discourse
RHETALHO (Pardo et al., 2004)	small	40	various	PT	discourse
Annodis (Afantenos et al., 2012)	75	156	various	FR	discourse
GloboQuotes (Fernandes et al., 2011)	1007	685	news	PT	quotes

Table 1: Overview of relevant resources annotated with some types of ARs or other relations overlapping with attribution.

swering (MPQA) opinion corpus (Wiebe, 2002), which annotates *private states*, such as beliefs and opinions, and the quotations introducing them; the corpora created for the Multilingual Opinion Analysis Task (MOAT) 7 and 8 (NTCIR) (Seki et al., 2008; Seki et al., 2010) which comprise opinion attributions to an explicit source; the Columbia Quoted Speech Attribution Corpus (CQSA) of narrative texts (Elson and McKeown, 2010). None of them is fully annotated with ARs. In the MPQA and NTCIR corpora, attribution is partly annotated, together with opinions and sentiments. While the annotation of cue and sources is included, the text span corresponding to the content is not annotated. The CQSA instead annotates direct quotations only and does not comprise the annotation of the AR cue. In the PDTB, discourse connectives and arguments are potential AR contents for which an *attribution span* including source and cue mention is usually annotated. Attributions are missing or incomplete when not fully matching explicit discourse relations. We can consider the AR that corresponds to the second paragraph of Ex. (1)[2]:

(1) The reports, attributed to the Colombian minister of economic development, said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags, the analyst said.

(HOWEVER) *These reports were later denied by a high Brazilian official, who said Brazil wasn't involved in any coffee discussions on quotas*, **the analyst said**.

(BUT) The Colombian minister was said to have referred to a letter that he said President Bush sent to Colombian President Virgilio Barco, and in which President Bush said it was possible to overcome obstacles to a new agreement. (wsj_0437)

The content span of this AR, is partially included in all three discourse relations below: the two implicit ones, having *however* and *but* as connectives, and the one with discourse connective *later*. In order to reconstruct the full AR from the annotation, it is necessary to take all three discourse relations into account and merge together the text spans attributed to 'the analyst said'.

[2]The content span is marked in italics, while the attribution span is marked in bold.

1. The reports said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags (Arg1)

   HOWEVER (Implicit connective)

   These reports were later denied by a high Brazilian official (Arg2)

2. The reports said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags (Arg1)

   LATER (Connective)

   These reports were denied by a high Brazilian official (Arg2)

3. *who said Brazil wasn't involved in any coffee discussions on quotas* (Arg1)

   BUT (Implicit connective)

   The Colombian minister was said to have referred to a letter that he said President Bush sent to Colombian President Virgilio Barco, and in which President Bush said it was possible to overcome obstacles to a new agreement (Arg2)

The example shows that there is no exact correspondence between ARs and discourse arguments and therefore some ARs are incompletely annotated or not annotated at all. This situation occurs when part of the AR content does not correspond to a discourse argument or when the whole AR is included in a discourse argument as in Arg1 of *But* (relation 3 above). The AR embedded in Arg1 ('**who said** *Brazil wasn't involved in any coffee discussions on quotas*') is just not annotated.

Since no available resource was fully satisfactory for the purpose of studying ARs, the goal of this project was to create a large and complete resource able to support a wide range of studies and the development of automatic extraction systems. The PDTB was chosen as the starting point to develop such attribution corpus, since it comprises a large number of ARs and the annotation is more compatible with the proposed approach to attribution.

3. PARC 3.0

3.1. Annotation Schema

The proposed annotation scheme was first presented in Pareti and Prodanof (2010) and used to create the ITAC corpus. It is grounded on the annotation in the PDTB as it considers a similar range of ARs, i.e. assertions, opinions, facts

and eventualities, and also identifies the relation as lexically anchored, i.e. the element that establishes the relation is a textual element. The annotation of attribution in the PDTB identifies two elements, similar to the treatment proposed by Bergler (1992) which marks a *matrix clause* and a subordinate or complement clause. While this approach constrains the matrix clause to contain a reporting verb and a source expressed as subject NP, the *attribution span* in the PDTB contains a broader range of elements expressing the source and the relation anchor. The more comprehensive approach adopted makes this corpus a more suitable starting point to study attribution and the wide range of structures that can express it (see the analysis in Sec. 4.).
The scheme I developed for attribution separately tags the three constitutive components of an AR, similar to the annotation of opinions in the MPQA, which marks the *text anchor*, the *source* and the *target* of an opinion. It also introduces an optional fourth element. The AR components are:

- Source span, i.e. the mention of the entity the content is attributed to.
- **Cue span**, i.e. the lexical anchor of the relation that expresses the source attitude towards the content.
- *Content span*, i.e. the attributed text.
- SUPPLEMENT SPAN, i.e. any additional element relevant to the interpretation of the AR, such as expressing information.

With respect to the PDTB annotation scheme, the modified scheme further classifies the *attribution span* into source and cue and introduces the supplement as a generic label for additional information that affects the AR, e.g. recipient or circumstantial information as in Ex.(2).

(2) *"Ideas are going over borders, and there's no SDI ideological weapon that can shoot them down,"* he **told** [A GROUP OF AMERICANS] [AT THE U.S. EMBASSY] [ON WEDNESDAY]. (WSJ_0093)

The original annotation scheme and subsequent modifications (Pareti, 2012) also proposed to annotate the set of features included in the PDTB (attribution type, source type, factuality, scopal polarity) and two additional ones: authorial stance and source attitude. These features are currently not included in the annotation of PARC 3.0.
Instead, two automatically derived features are included: *quote status* and *level of nesting*. The quote status identifies whether the content of an AR is a direct, indirect or mixed quotation by identifying whether the content span is partly, completely or not enclosed by quotation marks.
The *level of nesting* accounts for the depth of an attribution, i.e. whether the AR is nested into another AR. For each AR, the level of nesting can be reliably computed by counting the number of AR contents it is contained within, taking the text as the zero level. Nesting is a measure that impacts the reliability of the information conveyed. The information within the content span of an AR is at least second-hand (i.e. the author reports what someone else has expressed) and in case of nested ARs it can be third-hand or more. For all sources involved, their bias and credibility will affect whether we trust the AR they establish to be truthful and the conveyed information to be accurate. Nesting can be thought of as a distance measure, or the path the information went through to reach the text we are reading.
A nested AR inherits from the embedding one not only the source, but also its relation with the content, i.e. the attitude it holds towards it. In Ex. (3), the content of the nested AR 'she will come back' is affected by both sources (Mary and John) and their trustworthiness. Moreover, in Ex. (3a), the writer presents the attitude of the first-level source as uncertain and a belief, while in Ex. (3b) she presents it as factual and as constituting an assertion.

(3) a. John **doubts** [that Mary **said** *she will come back*].

b. John **announced** [that Mary **said** *she will come back*].

3.2. Corpus Development

A first corpus of over 9,800 ARs, PARC 1.0, was compiled from existing PDTB annotations that were reconstructed and further annotated semi-automatically. This version was used to conduct preliminary analysis of attribution. A revised version of it, PARC 2.0, was employed in experiments on the automatic extraction of quotation ARs (O'Keefe et al., 2012; Pareti et al., 2013; Almeida et al., 2014).
Although already a large resource for attribution, not all ARs are annotated in PARC 2.0. Any analysis based on the incomplete data was therefore only tentative since the annotated ARs were not a representative and balanced subset of all ARs in the corpus: their annotation was subordinate and dependent on that of discourse relations.
In addition, incomplete data was also detrimental for the development of supervised attribution extraction components which were confronted with the challenge of learning from positive instances and unlabelled data. While it is possible to overcome this issue, having a completely annotated resource is preferable.
The initial corpus was therefore further annotated with missing and nested ARs. The resulting corpus, PARC 3.0, includes 19,712 ARs and is divided into three sections corresponding to the WSJ corpus folders:

- Train: folders 00-22
- Development: folder 24
- Test: folder 23

The annotations originate from three distinct annotation phases:

1. PDTB derived: around half of the ARs are derived from the partial annotation in the PDTB. They were reconstructed and their 'attribution span' further annotated as 'source' and 'cue'. There are some annotation errors in the original annotation, in particular some incomplete content spans. These have not been corrected.

2. New annotation: annotation of all missing first-level ARs

3. Nested annotation: annotation of nested ARs in the development and test sections and folders 0-11 of the training section.

New annotations of first-level and nested ARs were added only to the 1,833 WSJ documents classified as news [3]. *News* is by far the largest genre in the WSJ corpus. PARC 3.0 annotation is in-line and encoded in XML.

3.3. Inter-Annotator Agreement

The new annotation was manually performed by three linguist annotators. Approximately 7% of PARC 3.0 news texts were double-annotated, which allowed to compute reliable inter-annotator agreement scores for the identification of ARs and for the selection of the spans corresponding to source, cue, content and supplement. Since the annotators were annotating different text spans, the agreement was calculated using the *agr* metric proposed in Wiebe et al. (2005). This computes the proportion of commonly annotated relations with respect to the overall relations identified by annotator *a* and annotator *b* respectively.

For the identification of an AR, the *agr* for each annotators pair varies from .74 to .82, while the overall *agr* is .79.

For the commonly identified ARs it is possible to compute the *agr* for the annotation of the spans corresponding to *source*, *cue*, *content* and *supplement*. Overlap results are calculated by taking the mean of the *agr* scores for each individual span.

Agreement results for the annotation of the spans corresponding to *source*, *cue*, *content* and *supplement* are reported in Table 2 and show that cues are almost always commonly identified with exact boundaries and source and content spans also have very high *agr*: .91 and .94 respectively.

Since for a large proportion of ARs no supplement was identified, the *agr* for the supplement span was calculated by taking into account only the ARs for which a supplement was identified. The score of .46 *agr* is rather low. However, the annotation of the supplement was included as exploratory of the kind of elements that would also be relevant for an AR and left underspecified in order to learn from the annotation instead of forcing it into a predefined direction.

Annotators	Cue	Source	Content	Supp
AB	1.00	0.90	0.95	0.67
BC	1.00	0.94	0.94	0.50
AC	0.99	0.88	0.95	0.30
Overall	1.00	0.91	0.94	0.46

Table 2: PARC 3.0 span selection overlap *agr* metrics for each pair of annotators and averaged to calculate the overall agreement.

[3] A list of WSJ documents per genre: `http://www.let.rug.nl/~bplank/metadata/genre_files_updated.html`

The inter-annotator agreement was also calculated, on a small set of texts, for the task of annotating nested ARs. The overall *agr* for this task was .70.

4. Attribution Analysis

Element	Occ.	%	Examples
NE	7894	40.0	*Mr. Greenspan*
noun	6053	30.7	*an official, analysts*
pronoun	3800	19.3	*they, his, I*
implicit	1510	7.7	NONE
wh. pronoun	250	1.3	*who, which, that*
determiner	173	0.5	*some, many at Lloyd's*
numeral	32	0.2	*the two, one in ten*

Table 3: Type of AR sources in PARC 3.0 (occurrence (Occ.) and percentage (%)).

The analysis reported in this section presents some of the key findings and the statistics computed on the whole PARC 3.0 corpus. These include PDTB derived ARs as well as the new first-level annotations and the nested ARs.

4.1. Source Span

The source is explicitly expressed in 92% of ARs. The remaining are cases where a passive structure, an adverbial cue (e.g. 'reportedly') or ellipsis of the subject in a coordinate or subordinate clause conceal the source. The vast majority of source spans consist of noun phrases, and over 83% of AR sources are noun phrases in subject position.

Concerning the common assumptions that sources correspond to named entities (NEs), the corpus shows that proper nouns are only a relative majority of sources (40%) as reported in Table 3, while a considerable number of sources are expressed by common nouns (30.7%), only in part referring to an NE. In particular, plural common nouns (e.g. 'lawyers', 'officials', 'people', 'nerds', 'libertarians' and 'enthusiasts') usually refer to categories of people and hardly ever to NEs. Another common type of sources is represented by pronouns (personal (19.3%) but also relative or who (1.3%), indefinite and demonstrative (0.9%) and pronominal cardinal numbers (0.2%), some of which will refer to NEs, while others not. In addition, 7.7% of ARs have an implicit source. Implicit sources are not only associated with passive attributional structures, but also impersonal constructions with the cue verb in the infinitive (Ex. (4)) or gerund form.

(4) Just **to** **say** *the distribution system is wrong* doesn't mean anything, [...] (wsj_0082)

4.2. Cue Span

The cue is usually expressed by a verb, but nouns, prepositions or prepositional groups, possessives and adjectives and also adverbials can also have this function. The adopted approach assumes that for each AR there is one and only one cue. Therefore there has to be a textual element expressing the relation for the relation to exist and if two cues connect a source-content pair, they establish two ARs. This is the

case in Ex. (5), where there are two ARs: a fact (A know B) and a belief (A believe B).

(5) Analysts **know** and **believe** *that the market is at a turning point*.

PARC 3.0 contains 527 attributional verb types (reporting, manner and other verbs), 40% of which occur a single time as an AR cue, thus relying on a pre-compiled list of verbs for attribution extraction, which is a common approach in the literature, is not a satisfactory solution. Moreover, while cues are mostly verbs, in 8% of ARs the cue is not a verb, thus focussing on verbs only would miss those relations (see Table 4).

Element	Occ.	%	Examples
verb	18k	92	*say, want, shrug*
noun	765	3.9	*announcement, idea*
prep. group	392	2.0	according to, in the eyes of
adjective	244	1.2	*is sure/skittish/aware*
preposition	81	0.4	*under, for, by, in, to*
punctuation	50	0.3	*colon, quotation mark*
adverbial	34	0.2	*admittedly, reportedly*

Table 4: Type of attributional cue in PARC 3.0 (occurrence (Occ.) and percentage (%)).

Lexical cues can occur alongside punctuation clues, such as quotation marks and colon. In those cases, as in Ex. (6), where punctuation clues are the only cues in the text, they take the role of AR cue.

(6) KIM**:** *I got home, let the dogs into the house and noticed some sounds above my head, as if someone were walking on the roof, or upstairs. [...]* (wsj_1778)

4.3. Content Span

While ARs are mostly identified at the intra-sentential level, the relation can cross sentence boundaries. The data contains 1,727 ARs spanning 2 to 27 sentences. Moreover, around 12% of AR contents are discontinuous. This is usually the case when the attribution span expressing the source and cue is in a parenthetical construction or when the content span continues in a contiguous sentence without any further clues being required as in Ex. (7).

(7) *"The Caterpillar people aren't too happy when they see their equipment used like that,"* **shrugs** Mr. George. *"They figure it's not a very good advert."* (wsj_1121)

Unlike the source, the content element cannot be implicit. However, it can be expressed by an anaphoric pronoun (e.g. the cataphoric content in Ex. (8)). In other cases, the content is not present but simply alluded (e.g. *He said the truth/ two words/ what he had to say*). Those apparent ARs are not annotated since the text span corresponding to the content is not present, not even anaphorically.

(8) Although Paribas **denies** *it*, analysts say the new bid in part simply reflects the continuing rivalry between France's two largest investment banking groups. (wsj_1319)

Contents can be expressed by virtually any syntactic structure, however most content spans correspond to a clausal element. Also relatively frequent, around 8% of the cases, is the content corresponding to one or more noun phrases (NP) as in Ex. (9).

(9) Even if the government **does see** *various "unmet needs,"* national service is not the way to meet them. (wsj_2407)

4.4. Nested ARs

Nested ARs are almost absent from the literature and their extraction has yet to be addressed, nonetheless, they are rather frequent, particularly in news. In PARC 3.0, there are 2,689 nested ARs. In order to correctly quantify their incidence, we have to consider only those texts that were annotated with nested ARs. In these texts, the percentage of nested ARs is over 20% as Table 5 reports. This translates to almost 1 in 4 first-level ARs carrying a nested AR within their content span. Nested ARs are very rarely direct, are mostly not assertions and have a larger proportion of pronominal and implicit sources.

Level of Nesting	PARC 3.0	Nested Sec.
1st	17016 (86.4)	9747 (79.7)
2nd	2526 (12.8)	2321 (19.0)
3rd	163 (0.8)	161 (1.3)

Table 5: Level of Nesting distribution in PARC 3.0. Occurrence (and percentage) of first-level and nested (2nd and 3rd level) ARs. Results are given for the complete PARC 3.0 and relative to the texts that were specifically annotated with nested ARs (i.e. news texts in folders 0-11, 23 and 24 in the corpus).

4.5. Quote Status

Direct, indirect and mixed ARs present various characteristics and complexity. While the content span of a direct AR is easily identified, that of a mixed AR has less clear boundaries and that of an indirect AR cannot be identified based on punctuation clues. Hence, the quote status of an attribution affects the complexity of the annotation and the success of an AR extraction system.

While the main focus of attribution extraction studies is on direct ARs, the quote status distribution in PARC 3.0, presented in Table 6, downsizes their relevance. Direct ARs account for just over 14% of all ARs, the same portion also corresponds to mixed, while 72% are indirect ARs. There is also a significant difference in distribution between nested and non-nested ARs. For nested ARs, the percentage of direct ones drops to just 1.7% and that of mixed to 10.6%.

Q-Status	Non-nested	Nested	All
Direct	2771 (16.2)	45 (1.7)	2816 (14.3)
Indirect	11.8k (69.3)	2361 (87.6)	14.1k (72.0)
Mixed	2464 (14.4)	286 (10.6)	2750 (14.0)

Table 6: Quote Status distribution in PARC 3.0. Occurrence (and percentage) of direct, indirect and mixed ARs.

Nested ARs are in fact mostly indirect, since direct reporting presupposes a verbatim of the original utterance, which becomes less likely, and credible, for nested ARs.

5. Applications

PARC 2.0, was employed in experiments on the automatic extraction of quotation. O'Keefe et al. (2012) and Almeida et al. (2014) used the corpus to develop supervised models of speaker attribution of direct quotations. The attribution as well as the extraction of all type of quotations, i.e. direct, indirect and mixed, was addressed in Pareti et al. (2013).
I further extended and developed these models to extract not only contents and sources of quotations, but complete ARs in Pareti (2015). The resulting system is a pipeline of different components that automatically extracts complete ARs, by identifying and linking source, cue and content of each AR.
The pipeline model is able to identify ARs reasonably well when using gold data to feed the different components, reaching 85% *precision* and 79% *recall* over strict matches (i.e. source, cue, content spans exactly identified). These results show the potential of the system, however they are an optimistic measure, since gold data would not be normally available. When run on predicted data, strict *precision* and *recall* drop to 78% and 65% respectively.
The system, trained on PARC 3.0, can identify and link source, cue and content spans of an AR with significantly higher precision and recall than traditional syntactic and rule-based approaches. This allows us to take news texts and automatically identify different types of ARs in it, whether opinions, quotations or other types. We can not only connect the attributed text to its source, but also know the textual anchor of the relation. This is a relevant element since it characterizes the relation by determining its type, factuality and evidential value and by carrying the source attitude and the authorial stance.

6. Conclusion and Future Work

Since a major drawback of the preliminary PARC versions was the data being only partially annotated, a second round of annotation was conducted in order to create a complete resource. This led to PARC 3.0, a corpus of almost 20k ARs. PARC 3.0 represents a large resource which comprises a broad range of ARs (e.g. quotations, opinions, facts and beliefs; direct, indirect and mixed; nested). The corpus, which is available for research use[4], lays the foundations for further studies on attribution extraction and can be used to develop supervised machine learning approaches.
Apart from enabling the development of extraction systems, PARC 3.0 has already allowed reaching a deeper understanding of the encoding of ARs in news. From the statistical analysis on the corpus and the results of the experiments on the extraction, we now know that:

- A significant proportion of ARs have no explicit source. While the quotation attribution literature starts from the assumption that all quotations have a source and address the task as a speaker attribution task, this approach is not suitable for a relatively small number of ARs.

- The majority of ARs are not delimited by quotation marks, thus their identification cannot be taken for granted. Identifying content spans and their boundaries for indirect and mixed ARs actually constitutes the hardest challenge for AR extraction.

- ARs are a more complex phenomenon than it appeared from the literature. They are not simply a syntactic phenomenon. This is clear just by considering that around 8% of ARs are inter-sentential. Therefore merely syntactic approaches to the extraction of ARs lead to systems that are relatively precise on a subset of ARs, but have rather low recall.

- Although disregarded by the literature, nested ARs are a large proportion of attributions in news, where even more than 20% of ARs may be nested. Nesting is not just a recursive aspect of attribution, this subset of the relation presents its own peculiarities and less typical encoding with respect to first-level ARs, making it the hardest type of ARs to identify.

- Attribution has been studied in different linguistic areas, however, there is no exact overlap of attribution for with of them. Attribution cannot be easily reduced to a syntactic or discourse phenomenon. It does show strong interconnections with other levels of linguistic analysis and it has important implications for factuality and opinion studies, however, it remains a separate task.

- Some of the assumptions at the basis of several approaches in the literature are not confirmed by the data; in particular, the assumptions that content spans are clausal elements, sources are NEs and cues are verbs. While these are frequent cases, the corpus shows that a relevant proportion of ARs does not fit these constraints.

While the current encoding of attribution is rather comprehensive, some additions would be desirable. In particular, it would be useful for the annotation to also encode the entity the source refers to. This would enable supporting entity resolution for the source, which is a crucial step for opinion and quotation attribution studies. For opinion studies it would be relevant to also annotate the target of the opinion attribution. Currently, this element is either included in the content span or marked as supplement, depending on how it is expressed. Another future addition should include the proposed features. Since different areas of study address different types of ARs, the *attribution type* would be a relevant aspect to add since it would allow to just select e.g. assertions or opinions. Moreover, it would be useful for factuality studies since the attribution type expresses the source's commitment towards the truth of the content and thus has implications on its factuality.

[4]There are plans to release PARC 3.0 through the LDC. At present, it can be obtained by contacting directly the author of this paper.

7. Acknowledgements

The creation of PARC 3.0 was possible thanks to the support of the Erasmus Placement programme and the Scottish Informatics and Computer Science Alliance (SICSA) Prize Studentship. Sincere thanks to all the annotators and to Prof. Bonnie Webber for her unfailing guidance.

8. Bibliographical References

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, M., Draoulec, A. L., Muller, P., Pery-Woodley, M.-P., Prevot, L., Rebeyrolles, J., Tanguy, L., Vergez-Couret, M., and Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Almeida, M. S. C., Almeida, M. B., and Martins, A. F. T. (2014). A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48, Gothenburg, Sweden, April. Association for Computational Linguistics.

Bergler, S. (1992). *Evidential analysis of reported speech*. Ph.D. thesis, Brandeis University, Waltham, MA, USA.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. Technical report ISITR- 545. Technical report, ISI, University of Southern California, Sept.

Elson, D. K. and McKeown, K. R. (2010). Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*.

Fernandes, W., Motta, E., and Milidiú, R. (2011). Quotation extraction for Portuguese. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011)*, pages 204–208, Cuiaba.

Li, H., Cheng, X., Adson, K., Kirshboim, T., and Xu, F. (2012). Annotating opinions in German political news. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

O'Keefe, T., Pareti, S., Curran, J., Koprinska, I., and Honnibal, M. (2012). A sequence labelling approach to quote attribution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Jeju, Korea.

Pardo, T. and Nunes, M. (2003). A construção de um corpus de textos científicos em português do Brasil e sua marcação retórica. Technical report, São Carlos-SP, September.

Pardo, T., das Graças Volpe Nunes, M., and Rino, L. (2004). Dizer: An automatic discourse analyzer for Brazilian Portuguese. In Ana Bazzan et al., editors, *Advances in Artificial Intelligence – SBIA 2004*, volume 3171 of *Lecture Notes in Computer Science*, pages 224–234. Springer, Berlin, Heidelberg.

Pareti, S. and Prodanof, I. (2010). Annotating attribution relations: Towards an Italian discourse treebank. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC*, Malta. European Language Resources Association (ELRA).

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. (2013). Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA, October. Association for Computational Linguistics.

Pareti, S. (2009). Towards a discourse resource for Italian: Developing an annotation schema for attribution. Master's thesis, Università degli Studi di Pavia, Pavia, 29 September.

Pareti, S. (2012). The independent encoding of attribution relations. In *Proceedings of the Eight Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-8)*, Pisa, Italy, October.

Pareti, S. (2015). *Attribution: A Computational Approach*. Ph.D. thesis, The University of Edinburgh, Institute for Language, Cognition and Computation, Edinburgh.

Pouliquen, B., Steinberger, R., and Best, C. (2007). Automatic detection of quotations in multilingual news. In *Proceedings of the International Conference Recent Advances In Natural Language Processing (RANLP 2007)*, pages 487–492.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation LREC*, pages 2961–2968.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003). The timebank corpus. In *Corpus linguistics*, pages 647–656.

Saurí, R. and Pustejovsky, J. (2009). Factbank: A corpus annotated with event factuality. *In Language Resources and Evaluation*, (43):227–268.

Seki, Y., Evans, D. K., Ku, L.-W., Sun, L., Chen, H.-H., and Kando, N. (2008). Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of NTCIR-7 Workshop Meeting on Evaluation of Information Access Technologies*, pages 185–203, NII, Japan, December.

Seki, Y., Ku, L.-W., and Sun, L. (2010). Overview of multilingual opinion analysis task at NTCIR- 8. In *Proceedings of NTCIR-8 Workshop Meeting on Evaluation of Information Access Technologies*, pages 209–220, Tokyo, Japan, June 15–18.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Wiebe, J. (2002). Instructions for annotating opinions in newspaper articles. Technical report, University of Pittsburgh, Pittsburgh, PA.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31:249–288, June.