A Database of Attribution Relations

Silvia Pareti

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

Abstract

The importance of attribution is becoming evident due to its relevance in particular for Opinion Analysis and Information Extraction applications. Attribution would allow to identify different perspectives on a given topic or retrieve the statements of a specific source of interest, but also to select more relevant and reliable information. However, the scarce and partial resources available to date to conduct attribution studies have determined that only a portion of attribution structures has been identified and addressed. This paper presents the collection and further annotation of a database of attribution relations from the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008). The aim is to build a large and complete resource that fills a key gap in the field and enables the training and testing of robust attribution extraction systems

Keywords: attribution, discourse, annotation

1. Introduction

In news reporting, *attribution* (indicating who has expressed some information, with what stance towards it) can affect up to 90% of the sentences in news text (Bergler et al., 2004). Aspects of attribution (its *source*, the stance, circumstantial information, etc.) can deeply affect how the information is received, as shown by comparing Ex.(1a) with the change in attitude in Ex.(1b) or the change in source in Ex.(1c).

(1) a. Dr. Smith **said**: "There is no correlation between smoking cigarettes and lung cancer."

b. Dr. Smith **jokes**: "There is no correlation between smoking cigarettes and lung cancer."

c. A smoker **said**: "There is no correlation between smoking cigarettes and lung cancer."

Being able to automatically identify the elements of an attribution relations (detailed below) would benefit sentiment analysis, summarisation and information retrieval tasks, e.g. enabling searches based on specific sources, presenting information from different perspectives, selecting more relevant and reliable results.

Despite several studies have addressed the automatic extraction of attribution, current systems such as NewsExplorer (Pouliquen et al., 2007) have been only partially successful. Failing to reach broad coverage (e.g. recall currently ranges from 13% (Pouliquen et al., 2007) to 66% (Wiegand and Klakow, 2010)) as well as satisfactory precision (with results in the 55-60% range (Schneider et al., 2010; Wiegand and Klakow, 2010)) these approaches are not robust enough to be reliably employed.

One reason for such weak performance is the lack of a comprehensive theory of attribution and of a complete annotated resource able to drive the development of extraction systems. Hence, studies have stalled as a consequence of false assumptions, thus addressing only some sub–parts of attribution and eluding the complexity of this relation.

In particular, a common approach has limited attention to verb cues and these, to only those attribution verbs commonly used in reporting (Krestel, 2007; Sarmento and Nunes, 2009). Another widespread assumption has limited the source of an attribution to be a NE (e.g. (Kim et al., 2007; Pouliquen et al., 2007)). This weakens performance whenever non–named entities and/or anaphoric NPs are involved.

Underestimating both the complexity of attribution, which presents considerable variation in structure and make-up, and the amount of resources needed to investigate attribution and test extraction systems, is responsible for our only partial success.

Starting from the partial annotation of attribution in the PDTB (Prasad et al., 2007; Prasad et al., 2008), this paper describes the construction of a large attribution database and presents data-driven analysis of how attribution relations are expressed. Section 2. briefly describes the complexity of this relation. Section 3. presents an overview of the resources available to date, along with their limitations. The collection of the attribution database and its further annotation are described in Section 4. together with the results from the agreement study evaluating the annotation schema. Section 5. contains the analysis of the collected data and Section 6. identifies the contribution of this study and the remaining issues to be addressed by the field.

2. Attribution

Attribution relations can be deconstructed as having three main elements:

- *content*, i.e. the attributed material
- source, i.e. the entity holding the content
- **cue**, i.e. the lexical anchor linking them

The effect of attribution is the insertion of a third party who "owns" the attributed material, i.e. an utterance (Ex.(2)), a belief or knowledge or an intention (Ex.(3)). The complexity of attribution is partly due to the variety of expressions

encoding it that makes the definition of a predictive structure not viable. Its *content* can range from a single word to multiple sentences (Ex.(4)). Its source can be expressed by a named (Ex.(4)) or unnamed (Ex.(2)) animate or inanimate entity, or it can be left implicit. Its **cue** can be a reporting verb (Ex.(2)), another verb (e.g. a manner verb as in Ex.(4)), a noun (Ex.(3)), an adjective, a preposition or adverb (e.g. according to, reportedly) or even just punctuation markers.

(2) Some members of the huge crowd **shouted** *"Viva peace, viva"*. (wsj_0559)[1]

(3) ...Mr. Lawson**'s promise** *that rates will be pushed higher if necessary*. (wsj_1500)

(4) "*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)

3. Resources

Attribution relations have been included in a small number of discourse annotation projects, such as the**RST Discourse Treebank** (Carlson and Marcu, 2001) and the **GraphBank** (Wolf and Gibson, 2005) (385 and 135 news articles respectively). The first annotates only intra–sentential attributions with an explicit source and a verb cue, the latter annotates attribution if no other discourse relation is present.

Discourse relations are also the focus of the annotation in the **PDTB** (Prasad et al., 2008), a collection of over 2,000 news articles from the WSJ. Attribution is not annotated as a discourse relation itself but rather as a feature of discourse relations and their arguments.

One of the most widely used resources for attribution-related studies is the MPQA **Opinion Corpus** (Wiebe, 2002), consisting of 692 documents (WSJ, American National Corpus, ...) annotating private states at the intra–sentential level. However, this resource has employed a sentence–based approach to attribution, limiting the attributions retrieved by those systems developed from this resource (e.g. Kim and Hovy (2005; Wiegand and Klakow (2010)).

The only corpus dedicated to the annotation of a wide range of attribution relations and their attributes is the **Italian Attribution Corpus** (ItAC)[2] (Pareti and Prodanof, 2010). It annotates *source*, *cue*, *content*, *supplement* and additional features. Despite its limited size (50 news articles), this corpus has allowed the identification of several attribution structures not addressed by the previous literature.

4. Attribution Database

4.1. Data Collection

The attribution database described here was collected starting from the annotation of attribution in the PDTB, where each discourse connective and its two arguments is associated with an attribution span of text where the attribution relation is established. Also annotated are source type (i.e. writer, other or arbitrary), attribution type (fact, e.g *know*, assertion, e.g. *say*, eventuality, e.g. *order*, belief, e.g. *think*), determinacy and scopal polarity, accounting for the factuality of the attribution itself.

Since the content of a newspaper article is by default attributed to its writer, unless otherwise expressed, such attribution relations have been excluded from the database. Attribution relations had to be reconstructed joining discourse connectives and arguments having the same attribution span into a same content span. The example in Fig. 1 reports the annotation in the PDTB of two discourse connective and relative arguments corresponding to the attribution relation in Ex.(5). The attribution span is reported in the *Text* field of the discourse connective, while the content of the attribution is fragmented, as it comprises the argument texts of both discourse connectives and the explicit discourse connective itself.

(5) *"There's no question that some of those workers and managers contracted asbestos–related diseases,"* **said** Darrell Phillips, vice president of human resources for Hollingsworth & Vose. *"But you have to recognise that these events took place 35 years ago. It has no bearing on our work force today."* (wsj_0003)

Each attribution relation was reconstructed, further annotated, as described in the next Section, and stored as stand–off annotation. The annotation includes, for each attribution, the elements showed in Table 1, together with reference to the original text for each annotated span.

```
____Explicit____
3904..3907
#### Text ####
But
#### Features ####
Ot, Comm, Null, Null
3820..3901
#### Text ####
said Darrell Phillips, vice president of
human resources for Hollingsworth &
Vose
####but, Comparison.Contrast
____Arg1____
3721..3817
#### Text ####
There's no question that some of those
workers and managers contracted
asbestos-related diseases
#### Features ####
Inh, Null, Null, Null
____Arg2____
3908..3971
#### Text ####
you have to recognize that these events
took place 35 years ago
#### Features ####
Inh, Null, Null, Null
```

```
____Implicit____
3973
#### Features ####
Ot, Comm, Null, Null
3820..3901
#### Text ####
said Darrell Phillips, vice president of
human
resources for Hollingsworth & Vose
####in other words, Expansion,
Contingency
____Arg1____
3930..3971
#### Text ####
that these events took place 35 years
ago
#### Features ####
Ot, Ftv, Null, Indet
3908..3929
#### Text ####
you have to recognize
____Arg2____
3973..4014
#### Text ####
It has no bearing on our work force
today
#### Features ####
Inh, Null, Null, Null
```

Figure 1: Annotation of attribution in the original release of the PDTB 2.0(Prasad et al., 2008). Each column reports the annotation relative to a discourse connective and its arguments, including its attribution.

[1]Examples are taken from the WSJ corpus. Sources are underlined, cues in bold and contents in italics.

[2]Freely available from:
http://homepages.inf.ed.ac.uk/s1052974/resources.php

ATTRIBUTION ID:	wsj_0003.pdtb_05
SOURCE SPAN:	Darrell Phillips, vice president of human resources for Hollingsworth & Vose
CUE SPAN:	said
CONTENT SPAN:	"There's no question that some of those workers and managers contracted asbestos–related diseases,"|"But you have to recognise that these events took place 35 years ago. It has no bearing on our work force today."
SUPPLEMENT SPAN:	None
FEATURES:	Ot, Ftv, Null, Indet
ATTRIBUTION STYLE:	Direct

Table 1: Example of an attribution relation entry in the attribution database.

RULE	EXAMPLE (WSJ)
(NP-SBJ)(VP)	one person **said**
(PP-LOC)(NP)(VB)	In Dallas, LTV **said**
(NP-SBJ)(VBP)(JJ)	I **am sure**

Table 2: Examples of matching rules for the annotation of the reporting span.

4.2. Further Annotation

The collected attribution relations have been further annotated in order to distinguish the elements in the reporting span. Around 80% of the annotation was performed semi–automatically, making use of a system of 48 rules (Table 2) to identify the most common source–cue patterns, and then manually corrected. The patterns specify lexical and syntactic features of source and cue elements in the span that match the rules, as well as additional elements relevant for the attribution. The remaining 20% of attribution spans, presenting less common structures, required manual annotation. This was performed by one expert annotator.

Elements of the attribution span have been marked as *source*, *cue* or *supplement*, according to the annotation schema developed in (Pareti and Prodanof, 2010). The source comprises the source mention together with its description, usually in the form of an appositive (Ex.(6)) or a relative clause. In case of a source expressed by a possessive adjective (Ex.(7)) or pronoun, the whole NP has been annotated.

(6) Pierre-Karl Peladeau, the founder's son and the executive in charge of the acquisition, **says** *Quebecor hasn't decided how it will finance its share of the purchase*, but he **says** *it most likely will use debt*. (wsj_0467)

(7) **His point**: *It will be increasingly difficult for the U.S. to cling to command-and-control measures if even the East Bloc steps to a different drummer.* (wsj_1284)

Verbal cues have been annotated together with their full verbal group, including auxiliaries, modals and negative particles. Adverbials adjacent to the cue, as in Ex.(8), have also been included, since they can modify the verb. Other parts of the verbal phrase have been marked as supplement. Prepositional cues (e.g. according to, for), adverbial cues (e.g. supposedly, allegedly), and noun cues (e.g. pledge, advice) have also been annotated.

(8) *"I'm not sure he's explained everything,"* Mrs. Stinnett **says grudgingly**. (wsj_0413)

All those elements that are also relevant for the interpretation of the content, but not strictly part of the attribution have been annotated as supplement. This includes circumstantial information, e.g. time (e.g. People familiar with Hilton said OVER THE WEEKEND (wsj_2443)), location, manner (e.g. Ex.(8)), topic (e.g. ON THE PROVISIONS OF THE MINNESOTA LAW, the Bush administration said ... (wsj_2449)) and recipient (He told THE WOMAN'S LAWYER, VICTOR BLAINE ... (wsj_0469).

Punctuation has also been added to the attribution database to distinguish between direct and indirect attributions, i.e. if the content of the attribution is expressed in quotes or not. The level of nesting of each attribution relation, i.e. an attribution being embedded into another attribution, has not been annotated since it can be automatically derived by identifying if each attribution is contained in the content span of another attribution.

4.3. Annotation Agreement

Inter-annotator agreement was not reported for the annotation of attribution in the PDTB, since it was done by a single annotator. The same holds for the ItAC corpus, a pilot application of the annotation schema adopted by the present study. To ensure the soundness of the annotation schema adopted, a portion of the WSJ has therefore been annotated by two expert annotators, following the instructions provided in the annotation manual. This makes use of examples as well as lexical clues to drive the annotation.

After familiarising themselves with the annotation guidelines and training themselves on a single article, the annotators independently annotated 14 articles with attribution relations and their features. The annotators were asked to identify an attribution relation and mark its constitutive elements (i.e. source, cue, content and supplement) and join them in a relation set. Subsequently, they would select values for the features: type (i.e. assertion, belief, fact or eventuality), source (i.e. other, arbitrary or writer), factuality (i.e. factual, non–factual) and scopal change (i.e. scopal change, none).

Overall the annotators identified 491 attribution relations in these 14 articles, with an agreement of 0.87. Since the spans annotated could differ, agreement was calculated with the *agr* metric (Wiebe et al., 2005). This metric represents the directional agreement of one annotator with re-

Feature	Cohen's Kappa
Type	0.64
Source	0.71
Scopal change	0.61
Factuality	0.73

Table 3: Kohen's Kappa values for agreement on the feature selection task.

spect to the other and it reflects the proportion of attribution relations identified by one annotator that were also identified by the other annotator. Several disagreements for this task were caused by one annotator including some instances of sentiments in the annotation, although these are outside the scope of this study.

Agr metrics were also calculated for each markable selection: Source, cue, content and supplement. The agreement on the selection of the span to annotate was high, with cue spans having an *agr* of 0.97, sources of 0.94 and contents of 0.95. Only for the supplement span there was little agreement (i.e. 0.37, calculated excluding the instances were both annotators marked no supplement). This element was however only optional and included to allow the annotation of additional relevant elements perceived as affecting the attribution.

The agreement for the feature selection was calculated using Cohen's Kappa. The results reported in Table 3 show rather low values. These could partly be improved with additional training and a simplification of the annotation effort, that should be separate in less complex sub–tasks. Agreement was also affected by the skewed distribution of features, with one value occurring in the majority of the instances (e.g. only 9 attributions were identified by one or both annotators as presenting a scopal change). The annotators were therefore confronted with only a limited number of instances for the less frequent feature values, and this could have led to more difficulty in recognising them and making consistent judgements.

5. Data Analysis

Preliminary analysis of the 9868 attributions in the database has identified some characteristics of attribution relations in news texts. In particular, while the majority of sources are indeed NEs (see Figure 2), their proportion has been overestimated in the literature. While most of the recent attribution extraction studies are concerned with attribution in news texts, early works have addressed narrative texts (Zhang et al., 2003) and their observations do not necessarily hold true for all types of text. Focusing on narrative, Elson and McKeown (2010) indicates sources as always being NEs, expressed by anaphoric pronouns in just about 9% of all cases. While this might be the case for narrative, in news texts a rather large proportion of sources are expressed by an anaphoric pronoun or a common noun, and excluding their extraction (as in Pouliquen et al. (2007)) is detrimental to performance.

Cues are expressed predominantly by verbs (96%) or *according to* (3%), while the remaining 1% comprises nouns, other prepositions, adverbials and punctuation–only cues.

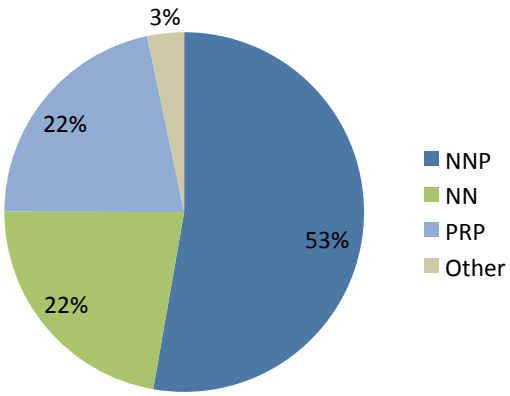

Figure 2: Source composition (NNP: proper noun, NN: common noun, PRP: pronoun).

Most extraction studies have considered only some reporting verbs as cues (only 35 such verbs used in Pouliquen et al. (2007)). It is fundamental to extend the list of verbs to comprise a larger number of reporting and non-reporting verbs. In the attribution database, *say* accounts for 70% of verb cues, followed by *add*, a non-reporting verb, *note*, *think* and *believe*. The remaining 261 verbs identified account for 20% of the verb cues (e.g. quip, smile). In order to identify the verbs most strongly associated with attribution, their predictivity was estimated based on their attributional versus overall usage in the original PDTB annotation of attribution (see Table 4). The upper limit here is less than 1, since (as noted earlier) the original annotation of attribution in the PDTB does not cover cases that were not annotated as a discourse relation or one of its arguments. By this estimate, the most predictive fifty verbs account in the PDTB corpus for about 90% of all attribution relations.

VERB	REPORTIVE	OVERALL	RATIO
say	6453	10643	0.60
quip	4	6	0.66
acknowledge	36	68	0.52
insist	27	11	0.24
continue	9	720	0.01

Table 4: PDTB attributive / overall verb occurrence ratio.

Concerning the content of attribution, this can be expressed in quotes (attribution of direct reported speech), not in quotes (indirect attributions) and partly in quotes (mixed attributions). *Direct* attributions can be more easily identified, making use of punctuation clues, and have been included in all studies while only some (e.g. Schneider et al. (2010) have addressed also *mixed* and *indirect* attributions). In the collected database, there are 2,290 *direct* attributions, (around 23% of all attribution relations), while the vast majority of attributions are *indirect* (5,920 instances) and a smaller proportion *mixed* (1,658 instances).

Although not included in the annotation, the level of nest-

ing of each attribution relation was automatically computed for the attributions in the annotation agreement study. Each attribution was assigned a number value according to its inclusion in the content of one or more other attributions. The results suggest that a very high proportion of attributions are nested in news language. Overall, 22% of the annotated attributions were nested, with a small proportion of attributions, about 3%, being nested in an already nested attribution.

6. Conclusion and Future Work

This paper motivates the need for a large resource annotated for attribution relations and describes the collection and further annotation of a database of over 9800 attributions from the PDTB. The resource was collected to fill the gap between the theory of attribution and current attribution extraction studies by building a resource that can be employed to develop and test broad–coverage attribution extraction systems.

The database can help deepen our understanding of attribution and verify intuitions based on occurring data. Preliminary data analysis has shed light on some of the unmotivated assumptions in the current literature (e.g. that sources are NEs and that a small set of reporting verbs can be sufficient to identify attribution) and identified some relevant aspects of attribution in news texts.

In the future, the database will be employed to investigate features of attribution affecting the content, such as different type of sources (e.g. anonymous, individuals, groups), authorial stance and source attitude.

Acknowledgements

This study was possible thanks to the contribution of Tim O'Keefe, a PhD student at The University of Sidney working on attribution, and the constructive supervision of Professor Bonnie Webber. The author is supported by a Scottish Informatics & Computer Science Alliance (SICSA) studentship.

7. References

Sabine Bergler, Monia Doandes, Christine Gerard, and René Witte. 2004. Attributions. In Yan Qu, James Shanahan, and Janyce Wiebe, editors, *Exploring Attitude and Affect in Text: Theories and Applications*, Technical Report SS-04-07, pages 16–19, Stanford, California, USA, March 22–25. AAAI Press. Papers from the 2004 AAAI Spring Symposium.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report ISITR- 545. Technical report, ISI, University of Southern California, September.

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

Soo-Min Kim and Eduard Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains, Pennsylvania, US*.

Youngho Kim, Yuchul Jung, and Sung-Hyon Myaeng. 2007. Identifying opinion holders in opinion text from online newspapers. In *Proceedings of the 2007 IEEE International Conference on Granular Computing*, GRC '07, pages 699–702, Washington, DC, USA. IEEE Computer Society.

Ralf Krestel. 2007. Automatic Analysis and Reasoning on Reported Speech in Newspaper Articles. Master's thesis, Universität Karlsruhe (TH), Fakultät für Informatik, Institut für Programmstrukturen und Datenorganisation (IPD), Karlsruhe, Germany.

Silvia Pareti and Irina Prodanof. 2010. Annotating attribution relations: Towards an Italian discourse treebank. In N. Calzolari et al., editor, *Proceedings of LREC10*. European Language Resources Association (ELRA).

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 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.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2007. Attribution and its annotation in the Penn Discourse TreeBank. *TAL (Traitement Automatique des Langues)*, 42(2).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation LREC08*.

Luis Sarmento and Sérgio Nunes. 2009. Automatic extraction of quotes and topics from news feeds. In *Proceedings of DSIE'09 - 4th Doctoral Symposium on Informatics Engineering*.

Nathan Schneider, Rebecca Hwa, Philip Gianfortoni, Dipanjan Das, Michael Heilman, Alan W. Black, Frederick L. Crabbe, and Noah A. Smith. 2010. Visualizing topical quotations over time to understand news discourse. Technical Report T.R. CMU-LTI-10-013, Carnegie Mellon University, Pittsburgh, PA, July.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Janyce Wiebe. 2002. Instructions for annotating opinions in newspaper articles. Technical report, University of Pittsburgh.

Michael Wiegand and Dietrich Klakow. 2010. Convolution kernels for opinion holder extraction. In *HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 795–803, Morristown, NJ, USA. Association for Computational Linguistics.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Comput. Linguist.*, 31:249–288, June.

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