Annotating Attribution Relations across Languages and Genres

Silvia Pareti
University of Edinburgh, UK
Google Inc.
s.pareti@sms.ed.ac.uk

1 Introduction

In Pareti (2012) I presented an approach to the annotation of attribution defining it as a relation intertwined albeit independent from other linguistic levels and phenomena. While a portion of this relation can be identified at the syntactic level (Skadhauge and Hardt, 2005) and part of it can overlap with the argument of discourse connectives (Prasad et al., 2006), attribution is best represented and annotated as a separate level.

The present work will present the results of an inter-annotator agreement study conducted in order to validate the annotation scheme described in previous work (Pareti and Prodanof, 2010). The scheme takes a lexicalised approach to attribution and is an extension and modification of the one adopted in the Penn Discourse TreeBank (PDTB) (Prasad et al., 2006). It comprises a set of elements, identified by the text spans expressing them, and a set of features.

Preliminary applications of the scheme to annotate attribution in different languages (English and Italian) and genres (news, spoken dialogues and mailing thread summaries) will also be presented and discussed.

2 Annotation Scheme Validation

This section describes an inter-annotator agreement study that was conducted in order to evaluate the applicability of the proposed annotation scheme before it was adopted to complete the annotation of the WSJ corpus. The study also verifies the validity of the PDTB derived corpus of attribution relations (ARs) before it was employed for the development and testing of quotation extraction and attribution studies (O'Keefe et al., 2012; Pareti et al., 2013; Almeida et al., 2014).

2.1 Annotation Scheme

The AR is defined as constituted by three main elements. The text span expressing each element is annotated and labelled as:

1. *Content*, i.e. what is attributed: this is usually a clause, but it can range from a single word up to several sentences. Content spans can be discontinuous (Ex.(1)).

2. **Source**, i.e. the entity the content is attributed to: a proper or common noun or a pronoun. The source is annotated together with its modifiers (i.e. adjectives, appositives, relative clauses). Sources might be left implicit, e.g. in case of passive or impersonal constructions.

3. <u>Cue</u>, i.e. the link expressing the relation: an attributional verb or, less frequently, a preposition, a noun, an adverb, an adjective or a punctuation mark. Modifiers of the cue, usually adverbs or negation particles, are also included in the cue span.

(1) *"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."*[1]

Optional information perceived as relevant for the interpretation of the AR, because of completing or contributing to its meaning, can be marked and joined in the relation as SUPPLEMENT. This element was introduced to allow the inclusion of circumstantial information as well as additional sources (informers) (e.g. John knows FROM MARY ...) or recipients (e.g. 'the restaurant manager told MS. LEVINE...'(wsj_1692).

The scheme comprises also six features that have been considered for inclusion into the scheme and were tested through an inter-annotator agreement study. Four features correspond to those included in the PDTB annotation: type (assertion, belief, fact, eventuality), source type (writer, other, arbitrary), determinacy or factuality (factual, non factual) and scopal polarity or scopal change. Two additional features are also included, since they are relevant aspect of an attribution and can affect how the content is perceived: authorial stance and source attitude.

The authorial stance reflects the authorial commitment towards the truth of the AR content, and it is the expression of the reporter's voice (Murphy, 2005) and her beliefs (Diab et al., 2009). The annotation distinguishes between neutral (e.g. say), committed (e.g. admit) or non–committed (e.g. lie, joke) authorial stance. The source attitude reflects whether a sentiment is associated with the attitude the source holds towards the content. The annotation scheme allows for five different values: positive (e.g. beam, hail, brag), negative (e.g. decry, fume, convict), tentative (e.g. believe, ponder, sense), neutral (e.g. report) or other.

2.2 Study Definition

In order to test the annotation scheme and identify problematic aspects, a preliminary inter–annotator agreement study was developed on a sample of the WSJ corpus. This sub–corpus consists of 14 articles, selected in order to present instances of all possible attribution types and feature values. Two experts annotators were independently asked to annotate the articles using the MMAX2 annotation tool (Müller and Strube, 2006), following the instructions provided in the annotation manual.

Since annotators were annotating different text spans, the agreement was calculated using the *agr* metric proposed in Wiebe et al. (2005). The *agr* metric is a directed agreement score that can be applied to relation identification tasks where the annotators do not choose between labels for a given annotation unit, but have to decide whether there is or not a relation and the scope of the text span that is part of it. For two given annotators *a* and *b* and the respective set of annotations *A* and *B* the annotators performed, the score returns the proportion of annotations *A* that were also identified by annotator *b*.

2.3 Inter-anotator Agreement Results

The annotators commonly identified 380 attributions out of the overall 491 ARs they annotated. For the AR identification task, the *agr* metric was 0.87. This value reflects the proportion of commonly annotated relations with respect to the overall relations identified by annotator *a* and annotator *b* respectively (i.e. the arithmetic mean of $agr(a||b)$ 0.94 and $agr(b||a)$ 0.80). Higher disagreement correlated with the identification of nested attributions, i.e. ARs that appear within the content span of another AR. If overall 22% of the ARs identified by the annotators were nested, the proportion dropped to 15.5% for the ARs identified by both annotators. Nested ARs represent instead over 44% of the ARs identified only by one annotator.

The agreement with respect to choosing the same boundaries for the text spans to annotate was also evaluated with the *agr* metric. The results (Table 1) are very satisfactory concerning the selection of the spans for the source (.94 *agr*), cue (.97 *agr*) and content (.95 *agr*) elements. Concerning the supplement, there was instead little agreement as to what was relevant to the AR in addition to source, cue and content.

[1] Examples in this paper mark the source span of an attribution in **bold**, the content span in *italics* and the cue span as underlined.

Cue	Source	Content	Supplement
0.97	0.94	0.95	0.37

Table 1: Span selection *agr* metrics.

Features	Raw Agreement	Cohen's Kappa	N Disagreements
Type	0.83	0.64	63
Source	0.95	0.71	19
Scopal change	0.98	0.61	5
Authorial stance	0.94	0.20	21
Source attitude	0.82	0.48	67
Factuality	0.97	0.73	9

Table 2: Raw and Kappa agreement for the feature value selection.

Once having identified an attribution, the annotators were asked to select the values for each of the 6 annotated features. Several issues emerged from this task. Despite very high raw agreement values, the corrected Kappa measure shows a very different picture and results mostly below satisfactory. Only the selection of the source type and the factuality value are above the 0.67 recognised by some literature as the threshold allowing for some tentative conclusions, as discussed in detail by Artstein and Poesio (2008).

2.4 Agreement Discussion

The results of the agreement study allowed to identify some issues concerning the proposed features. In particular, the need for a better definition of the boundaries of each feature value. One of the difficulty in applying the proposed annotation schema originated from the number of elements and features that needed to be considered for the annotation of each attribution. This suggests that by decreasing its complexity, the number of errors could be reduced. The annotation should be therefore split into two separate task: the AR annotation and the feature selection.

For certain decisions, test questions could be a useful strategy to ensure a better convergence of the results, e.g. to determine whether the scope of a negation affects the content (and should be annotated as a scopal change) instead of the AR itself (thus affecting its factuality).

While a redefinition of some of the features and a simplification of the task would help reduce ambiguity, subjectivity and errors, the low agreement is also greatly affected by the imbalanced data. Most features assume one value in the majority of the cases, while some values appear only rarely. This has a detrimental effect on the annotator's concentration and ability to recognise these cases.

It is highly desirable to build a complete resource for attribution studies enriched by relevant features that affect the interpretation and perception of ARs. However, in the light of the inter-annotator agreement study, it was decided to restrict further annotation efforts to the AR span selection and postpone the annotation of the features.

3 Attribution in Italian and English

The scheme for the annotation of ARs was initially applied to Italian news articles, leading to the creation of a pilot corpus of 50 texts, the Italian Attribution Corpus (ItAC) (Pareti and Prodanof, 2010).

Attribution relations in Italian are expressed in a similar way as they are in English, thus the same scheme could be used for both languages. Unlike Italian, however, English can express attribution, to an unspecified source, by means of adverbials (e.g. reportedly, allegedly). These cases nonetheless fit the schema (see Ex.(2)) since sources can be left implicit.

(2) *Olivetti* reportedly *began shipping these tools in 1984.*

Table 3 shows a comparison of the Italian pilot (ItAC) and the English PARC 3.0 AR corpora. Both corpora were annotated with the scheme developed for attribution. Although very different in size, some patterns already emerge. The comparison shows a smaller incidence of ARs per thousand tokens in the Italian corpus. This is more likely due to differences in style between the news corpora or to cultural differences rather than to characteristics of the language.

A much higher proportions of ARs in Italian (around 29%) do not have an associated source span. The proportion of ARs without a source is in English rather small (8%) and mostly due to passive constructions and other expressions concealing the source. These cases have usually been disregarded by attribution extraction studies focusing on the identification of the entity the source refers to, since they do not refer to a specific entity or they refer to an entity that is not possible to identify.

Italian however is a pro-drop language, that is, subject pronoun are usually dropped since a rich verb morphology already includes person-number information and they are therefore superfluous. If we also consider that in PARC 3.0 over 19% of source mentions are pronouns, we can understand why Italian has around 20% more ARs without an explicit source than English. Unlike impersonal or missing AR sources in English, pro-drop sources in Italian usually refer to an entity and should be resolved.

	ItAC	PARC 3.0
Texts	50	2,280
Tokens	37k	1,139k
Toks/Text	740	500
ARs	461	19,712
ARs/text	9.2	8.6
ARs/1k tokens	12.5	17.3
ARs no source	29%	8%

Table 3: Comparison of AR news corpora of Italian (ItAC) and English (PARC 3.0) annotated with the AR scheme described in this work.

Some differences between the two languages emerged also concerning the choice and distribution of verbal cues. In a study comparing attribution in English and Italian opinion articles, Murphy (2005) noted that English commentators used more argumentative and debate seeking verbs while the Italian ones are more authoritative and consensus seeking. By looking at the verb type distribution in the two corpora, it is worth noting the high proportion of attributional ‘say’ in English, around 50% of all cue verbs, which has no parallel in Italian. This might have to do with a tendency towards using a more neutral language in English as well as with the Italian distaste for repetitions and the use of broad meaning verbs, considered as less educated.

The annotation scheme for attribution could be successfully applied to both English and Italian, since they do not present major differences in the structures they use to express attribution.

Other languages, however, can also express attribution morphologically, e.g. some agglutinative languages like Japanese, Korean and Turkish express reportative evidentiality with verb suffixes and particles. These languages would require more investigation to determine whether adaptations to the annotation scheme are necessary.

4 Cross-genre Applications

While extremely frequent and relevant in news, attribution is not a prerogative of this genre. Very little work exists addressing attribution in other genres and it is almost exclusively limited to narrative. PARC 3.0 already contains texts from different genres, albeit all related to news language. The WSJ files included in the PDTB have been classified into 5 different genres: essays, highlights, letters, errata and news. But what if we try to encode attribution in more distant genres and we take into account different registers and domains? In order to test this, I will present here two preliminary studies we developed, annotating attribution on very different kind of corpora: technical mailing thread summaries and informal telephone spoken dialogues.

	PARC 3.0	SARC	KT-pilot
Genre	News	Dialogue	Thread summaries
Register	Formal	Informal	Informal
Medium	Written	Oral	Written
Tokens	1,139k	16k,2h	75k
ARs	19,712	223	1,766
ARs/1k tokens	9.2	14	23

Table 4: Comparison of AR corpora from different genres annotated with the AR scheme described in this work.

4.1 Attribution in Mailing Thread Summaries

The annotation schema for attribution was applied by Bracchi (2014) to a pilot corpus of mailing thread summaries (KT-pilot) sampled from the Kernel Traffic Summaries of the Linux Kernel Mailing List[2]. The corpus differs not only in genre, but also in register and domain. The summaries report what different people contributed in writing to the discussion. This consists in a back and forth of comments and replies. The register is rather informal and the domain is technical. This corpus is particularly interesting for attribution since it is distant from the news genre, but it is also extremely rich in ARs. The corpus was studied by Duboue (2012), who investigated the varied ways of reporting that could be used in summaries.

While the schema was suitable to encode ARs in this genre, some differences emerged with respect to news texts. Bracchi (2014) reports preliminary analysis concerning the attribution cues. She identifies some characteristics of ARs cues in the KT-pilot, for example the use of acronyms as attribution spans, representing both the source and the cue (e.g. IMHO: 'in my humble opinion', AFAIK:'as far as I know', IMNSHO: 'in my not so humble opinion'). Since the annotation allows for the source and cue element to overlap, these cases can be annotated with the acronym corresponding both to the source and the cue span.

(3) *This* **IMHO** *is a good thing for all Real Time SMP.* (Bracchi, 2014)

As Bracchi (2014) notes, the occurrence of attributional verb cues in the KT-pilot is also more distributed, with 'say' covering only 18% of the cases (compared to around 50% in PARC 3.0) and almost 11% being covered by 'reply', a common verb in the mailing thread summaries but rather low-frequency in news. Moreover, some common verbs, strongly associated with attribution in news language (e.g. declare and support) exhibit in the computer domain of the KT-pilot a preferred other use (e.g. 'declare a variable', 'support a version').

4.2 Attribution in Spoken Dialogues

In Cervone et al. (2014), we investigated attribution in spoken informal telephone dialogues and explored the possibility to apply the proposed annotation scheme to a genre using a different medium of communication. The preliminary corpus (Speech Attribution Relation Corpus (SARC)) was annotated with a modification of the scheme for attribution. The basic scheme, with source, cue and content elements being annotated, could be applied to the dialogues, with the only addition of the 'fading out' category. This category is borrowed from Bolden (2004) to account for additional words whose inclusion in the content is ambiguous. In (4) the part of the content span delimited by square brackets is considered as fading out, since it is uncertain whether it still is part of what was originally uttered.

[2] (http://kt.earth.li/kernel-traffic/archives.html)

(4) **I** told him *that I cared a lot about him [because I mean I've always been there for him haven't I]*

Although typical of the spoken medium, where only the beginning of a source shift is signalled, 'fading out' has a parallel in written texts, where syntactic ambiguities can leave the content boundaries unclear as in the bracketed portion of the content in Ex.(5) which could be part of what the workers described as well as a remark the author adds. In PARC 3.0, it was up to the annotators to determine the boundaries of the content for each case, although indication was given as to adopt a minimal approach, thus excluding the ambiguous parts.

(5) **Workers** described *"clouds of blue dust" that hung over parts of the factory*, [even though exhaust fans ventilated the area].

Similarly to news, where the article attribution to its writer is not annotated, in SARC the relation between the speaker and each turn utterance in the dialogue is not annotated as an AR. While a dialogue in fiction or an interview in news articles would be an AR, turns in spoken dialogues are not. The turns are not linguistically expressed, as it is obvious to the participant in a spoken conversation what is uttered by a certain speaker (recognised by the voice or because we can see her speaking or because he is simply the other, the voice on the other side of the phone). The attribution of the text itself is not annotated since it is a meta-textual or extra-textual attribution. SARC annotates instead the ARs within a turn utterance.

Some smaller differences with respect to news derive from SARC being a corpus of spoken and colloquial language. Apart from the use of colloquial attributional expressions such as 'I'm like' or 'she goes' that are not likely to appear in news, there are frequent repetitions and broken sentences. In Ex.(6), the source and cue of the AR are repeated twice. In news language this would normally be a case of nested ARs (i.e. Ellie just said to me yesterday: "She said: 'Oh I'm a bit bored of the snow now mum'"). However, here there is only one AR and only the closest source and cue should be annotated since an AR should have only one cue. Each cue established a different AR (e.g. He thinks and knows that ...) although holding between the same source-content pair. While an AR can have multiple sources, this is intended to represent the case when a content is attributed to more than one source (e.g. 'toy manufacturers and other industrialists') and not twice to the same source.

(6) haven't ye ah God do you know I was just off it now and **Ellie** just said to me yesterday **she** said *oh I'm a bit bored of the snow now mum*

The application of a lexicalized approach to attribution to the spoken medium, proved more problematic. In particular, speech lacks punctuation, which instead plays a crucial role in written texts, allowing the identification of direct quotations and in some cases being the only lexical cue of an AR. In speech dialogues instead, part of the role played by punctuation is taken over by acoustic features. The preliminary analysis reported by Cervone (2014) shows some correlation of acoustic aspects, such as pauses, intensity and pitch, with the content boundaries. In the examples below (Cervone (2014)[p.102]), acoustic features allow to reconstruct the ARs in the dialogue turn in Ex.7a as it is shown in Ex.7b with the help of punctuation.

Moreover, not only the content boundary has to rely on extra-textual clues, but in certain cases, the whole AR is reduced in the text to its content element. In spoken language, cues might be expressed by acoustic features and thus not identifiable from the text alone. In the example (Ex.(7)), "what for a loft" and "I'm not going to do that" are attributed to a different source (mentioned at the beginning of the turn as 'she'). However, the source is left implicit and the cue replaced by acoustic means.

(7) a. she wouldn't I said well but I said at the end of the day I said you could sell your house what for a loft and I said well yes if you really didn't have any money you'd have to sell it for a loft buy something smaller well I'm not going to do that and I thought well then you haven't not got any money then have you it's not really the same thing

b. She wouldn't. I said: "Well but", I said: "At the end of the day", I said: "You could sell your house." "What? For a loft?" And I said: "Well, yes! If you really didn't have any money you'd have to sell it for a loft. Buy something smaller." "Well I'm not going to do that." And I thought: "Well, then you haven't not got any money then, have you?" It's not really the same thing.

4.3 Other Forms of Attribution

Not only in the spoken medium, but also in the web one, attribution can also be expressed in extra-textual ways, thus requiring a partly different encoding. For example, attribution can rely on hypertext, both to express the source and to delimit the content span by embedding in it a link to its source.

In addition, the web can make use of graphical elements to show the source of some text, e.g. by embedding part of another page or showing a tweet as an image. Attribution is also graphically expressed in the comics medium, where sources are drawn and cues are rendered by bubbles enclosing the text and encoding the type of attitude by means of specific shapes and by varying the line thickness or continuity.

Also in academic writing, attribution is expressed in a distinct way, with sources being papers commonly referenced in a strictly encoded way.

5 Conclusion

This papers discusses the validity and applicability of the annotation scheme for attribution relations that was proposed in previous work and adopted to annotate PARC 3.0, a large corpus of attribution built on the WSJ corpus. The scheme was tested with a small inter-annotator agreement study. The results showed relatively high agreement for the identification of an AR and very high agreement, over 90%, for the selection of the source, cue and content spans. On the other hand, there was little agreement for the selection of the attribution features, which suggests that they should be redefined and further tested before being included in the annotation.

The scheme has been applied both to English and Italian news corpora. While some differences between the two languages emerged, in particular the higher incidence of ARs with implicit source in Italian, attribution could be annotated in both languages without modifications to the scheme.

The paper reviews two additional pilot corpora from different genres. These were annotated with ARs in order to test the way attribution is expressed in genres other than news articles. While no substantial differences emerged when annotating mailing list thread summaries, the annotation of informal spoken dialogues posed more challenges. In speech, we identified the presence of acoustic elements reinforcing or even replacing the source and cue of an AR, thus showing that an approach solely based on lexical features is not viable for this genre.

Overall, preliminary applications of the current annotation scheme beyond English news texts showed good flexibility and coverage of the current approach. Nonetheless, in specific cases, some adaptation to different language structures and to different genres would be needed.

References

Almeida, M. S. C., M. B. Almeida, and A. F. T. Martins (2014, April). 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, Gothenburg, Sweden, pp. 39–48. Association for Computational Linguistics.

Artstein, R. and M. Poesio (2008, December). Inter-coder agreement for computational linguistics. Comput. Linguist. 34, 555–596.

Bolden, G. (2004). The quote and beyond: defining boundaries of reported speech in conversational russian. Journal of pragmatics 36(6), 1071–1118.

Bracchi, A. (2014, February). Attribution relation cues across genres: A comparison of verbal and non-verbal cues in news and thread summaries. In The 41st Language at Edinburgh Lunch, Edinburgh, UK, pp. poster.

Cervone, A. (2014). Attribution relations extraction in speech: A lexical-prosodic approach. Master's thesis, Università degli Studi di Pavia, Pavia.

Cervone, A., S. Pareti, P. Bell, I. Prodanof, and T. Caselli (2014, December). Detecting attribution relations in speech. In B. M. e. Roberto Basili, Alessandro Lenci (Ed.), First Italian Conference on Computational Linguistics CLiC-it 2014, Pisa, Italy, pp. poster. Pisa University Press.

Diab, M., L. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo (2009). Committed belief annotation and tagging. In Proceedings of the Third Linguistic Annotation Workshop, pp. 68–73.

Duboue, P. (2012). Extractive email thread summarization: Can we do better than he said she said? INLG 2012, 85.

Müller, C. and M. Strube (2006). Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Mukherjee (Eds.), Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pp. 197–214. Germany: Peter Lang.

Murphy, A. C. (2005). Markers of attribution in English and Italian opinion articles: A comparative corpus-based study. ICAME Journal 29, 131–150.

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

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

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

Pareti, S. and I. Prodanof (2010). Annotating attribution relations: Towards an Italian discourse treebank. In N. C. et al. (Ed.), Proceedings of LREC10. European Language Resources Association (ELRA).

Prasad, R., N. Dinesh, A. Lee, A. Joshi, and B. Webber (2006). Annotating attribution in the Penn Discourse TreeBank. In Proceedings of the Workshop on Sentiment and Subjectivity in Text, SST '06, pp. 31–38.

Skadhauge, P. R. and D. Hardt (2005). Syntactic identification of attribution in the RST treebank. In Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora.

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