10.12871/CLICIT2014121

Detecting Attribution Relations in Speech: a Corpus Study

Alessandra Cervone[1], Silvia Pareti[23], Peter Bell[2], Irina Prodanof[1], Tommaso Caselli[4]
Dept. of Humanities, University of Pavia[1], School of Informatics, University of Edinburgh[2], Google Inc.[3], Trento RISE[4]
alessandra.cervone01@universitadipavia.it; s.pareti@sms.ed.ac.uk; peter.bell@ed.ac.uk; irina.prodanof@unipv.it; t.caselli@trentorise.eu

Abstract

English. In this work we present a methodology for the annotation of Attribution Relations (ARs) in speech which we apply to create a pilot corpus of spoken informal dialogues. This represents the first step towards the creation of a resource for the analysis of ARs in speech and the development of automatic extraction systems. Despite its relevance for speech recognition systems and spoken language understanding, the relation holding between quotations and opinions and their source has been studied and extracted only in written corpora, characterized by a formal register (news, literature, scientific articles). The shift to the informal register and to a spoken corpus widens our view of this relation and poses new challenges. Our hypothesis is that the decreased reliability of the linguistic cues found for written corpora in the fragmented structure of speech could be overcome by including prosodic clues in the system. The analysis of SARC confirms the hypothesis showing the crucial role played by the acoustic level in providing the missing lexical clues.

Italiano. *In questo lavoro viene presentata una metodologia di annotazione delle Relazioni di Attribuzione nel parlato utilizzata per creare un corpus pilota di dialoghi parlati informali. Ció rappresenta il primo passo verso la creazione di una risorsa per l'analisi delle ARs nel parlato e lo sviluppo di sistemi di estrazione automatica. Nonostante la sua rilevanza per i sistemi di riconoscimento e comprensione del parlato, la relazione esistente tra le citazioni e le opinioni e la loro fonte è stata studiata ed estratta soltanto in corpora scritti, caratterizzati da un registro formale (articoli di giornale, letteratura, articoli scientifici). Lo studio di un corpus parlato, caratterizzato da un registro informale, amplia la nostra visione di questa relazione e pone nuove sfide. La nostra ipotesi é che la ridotta affidabilitá degli indizi linguistici trovati per lo scritto nella struttura frammentata del parlato potrebbe essere superata includendo indizi prosodici nel sistema. L'analisi di SARC conferma quest'ipotesi mostrando il ruolo cruciale interpretato dal livello acustico nel fornire gli indizi lessicali mancanti.*

1 Introduction

Our everyday conversations are populated by other people's words, thoughts and opinions. Detecting quotations in speech represents the key to "one of the most widespread and fundamental topics of human speech" (Bakhtin, 1981, p. 337).

A system able to automatically extract a quotation and attribute it to its truthful author from speech would be crucial for many applications. Besides Information Extraction systems aimed at processing spoken documents, it could be useful for Speaker Identification systems, (e.g. the strategy of emulating the voice of the reported speaker in quotations could be misunderstood by the system as a change of speaker). Furthermore, attribution extraction could also improve the performance of Dialogue parsing, Named-Entity Recognition and Speech Synthesis tools. On a more basic level, recognizing citations from speech could be useful for sentence boundaries automatic detection systems, where quotations, being sentences embedded in other sentences, could be a source of confusion.

So far, however, attribution extraction systems have been developed only for written corpora.

Extracting the text span corresponding to quotations and opinions and ascribing it to their proper source within a text means to reconstruct the Attribution Relations (ARs, henceforth) holding between three constitutive elements (following Pareti (2012)):

- the Source
- the Cue, i.e. the lexical anchor of the AR (e.g. *say, announce, idea*)
- the Content

(1) This morning [$_{\text{Source}}$ John] [$_{\text{Cue}}$ told] me: [$_{\text{Content}}$ "It's important to support our leader. I trust him."].

In the past few years ARs extraction has attracted growing attention in NLP for its many potential applications (e.g.Information Extraction, Opinion Mining) while remaining an open challenge. Automatically identifying ARs from a text is a complex task, in particular due to the wide range of syntactic structures that the relation can assume and the lack of a dedicated encoding in the language. While the content boundaries of a direct quotation are explicitly marked by quotation markers, opinions and indirect quotations only partially have syntactically, albeit blurred, boundaries as they can span intersententially. The subtask of identifying the presence of an AR could be tackled with more success by exploiting the presence of the cue as a lexical anchor establishing the links to source and content spans. For this reason, cues are the starting point or a fundamental feature of extraction systems (Pareti et al., 2013; Sarmento and Nunes, 2009; Krestel, 2007).

In our previous work (Pareti and Prodanof, 2010; Pareti, 2012), starting from a flexible and comprehensive definition (Pareti and Prodanof, 2010, p. 3566) of AR, we created an annotation scheme which has been used to build the first large annotated resource for attribution, the Penn Attribution Relations Corpus (PARC)[1], a corpus of news articles.

In order to address the issue of detecting ARs in speech, we started from the theoretical and annotation framework from PARC to create a comparable resource. Section 2 explains the issues connected with extracting ARs from speech. Section 3 describes the Speech Attribution Relations Corpus (SARC, henceforth) and its annotation scheme. The analysis of the corpus is presented in Section 4. Section 5 reports an example of how prosodic cues can be crucial to identify ARs in speech. Finally, Section 6 draws on the conclusions and discusses future work.

[1]The corpus adds to and further completes the annotation of attribution in the PDTB (Prasad et al., 2008).

2 The challenge of detecting Attribution Relations in speech

The shift from written to spoken language makes the task of ARs extraction much harder. Current approaches to ARs detection rely heavily on lexical cues and punctuation to identify ARs and in particular the Content span boundaries. In the fragmented structures full of disfluencies typical of speech, however, lexical cues become less reliable, sometimes being completely absent.

On the other hand, punctuation, in most cases crucial in giving the key to the correct interpretation of ARs, is replaced in speech by prosody. While punctuation is a formal symbolic system, prosody is a continuous system which could greatly vary due to language-specific, diatopic, diaphasic and idiosyncratic reasons, thus much harder to process for a tool.

Our working hypothesis focused on the role of prosody in marking the presence and boundaries of quotations in speech. In particular, we considered that it would be possible to find acoustic cues to integrate the linguistic ones in order to improve the task of correctly reconstructing the ARs in a spoken corpus.

Preliminary support for our hypothesis can be found in previous studies which aimed at identifying acoustic correlates of reported speech. However, these approaches, which suggest shift in pitch, intensity and pauses duration as possible prosodic indicators of quotations, offer only fragmented insights on the phenomenon of Attribution. Some of these studies analyze only the variations in pitch (Jansen et al., 2001; Bertrand et al., 2002), others analyze only the ending boundary of quotations (Oliveira and Cunha, 2004) and most of them consider only direct reported speech (Bertrand et al., 2002; Oliveira and Cunha, 2004). Even if the results of these studies are encouraging, the acoustic cues they propose need to be tested and further investigated in a larger project which consider different types of reported speech along with all the prosodic features which could be linked to quotations (pitch, intensity and pauses).

3 Description of the corpus

SARC is composed by four informal telephone conversations between English speakers. The dialogues have a mean duration of 15 minutes where every speaker is recorded on a different track (totally about 2 hours of recordings and 8 speakers). Table 1 shows the main aspects which differentiate SARC from PARC.

	SARC	**PARC**
Register	Informal	Formal
Medium	Oral	Written
Genre	Dialogue	News
Tokens	16k, 2h	1139k
ARs Frequency (ARs per k tokens)	(223/16k) 13.9	(10k/1139k) 9.2

Table 1: Differences between SARC and PARC.

While PARC displays a rather formal English register, typical of the news genre and of the written medium, SARC portrays a radically different one, the coloured, fragmented and highly contextualized register used in informal conversations. The impact of these differences in the type of language presented in our corpus have lead to an adaptation, summarized in Table 2, of the annotation scheme created for PARC (Pareti, 2012).

Attribution Elements	Source	
	Cue	
	Content	*Direct*
		Indirect
		Fading Out
Relation		

Table 2: Annotation scheme for SARC.

All the basic annotation elements (source, cue, content) from PARC have been kept in order for the results to be comparable. The content has been further subdivided into 3 types, of which the last one, the Fading out, never used previously in attribution extraction schemes, is a category introduced by Bolden (2004) to identify those cases typical of dialogues in which the ending boundary of a quotation is left purposely ambiguous by the speaker. We adopted PARC (Pareti, 2012; Pareti, forthcoming) annotation guidelines, with the following modifications: in PARC cases like *"I say"* or *"I think"* are considered quotations of the author himself, while in our annotation, where every sentence is considered a personal opinion of the speaker, they are not (see Klewitz and Couper-Kuhlen(1999, p. 4)). The annotation has been performed with MMAX2 (Müller and Strube, 2006) by one annotator (who was also trained on PARC scheme and guidelines). For further details about the construction and annotation process of SARC we refer you to Cervone (2014).

4 Analysis of SARC

The analysis of SARC (see the chart in Figure 1) shows how in about 10% of the cases in our corpus the cue is completely missing, while in PARC such cases were rare (only in 4% of the cases was the source missing). Therefore, at least 1 out of 10 ARs in SARC is impossible to identify without the aid of prosodic cues. Furthermore, due to

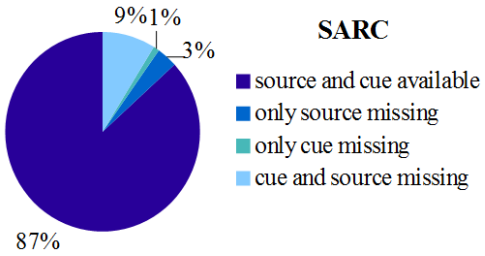

Figure 1: Cases of missing AR elements in SARC.

the absence of punctuation all the boundary clues found for written corpora are missing. We cannot rely any more on quotation marks, without punctuation we have no clue about the sentence structure (crucial for indirect quotes) and due to disfluencies the syntactic structure is less reliable and complete (some ARs are syntactically encoded). This means that even that 87% of cases in which in SARC no element of the AR is missing are more problematic than almost 50% of PARC cases (3,262 direct, 1,549 mixed) where punctuation defines the content. If we rely only on the lexical level for detecting ARs in speech, we have no assurance that the boundaries of the content span we identified out of many possible interpretations are the correct ones.

5 Prosodic cues of Attribution

The analysis of SARC has shown how much the shift to a spoken corpus can make the task of detecting ARs harder, displaying the need to find other cues to improve the performance of an attribution extraction system for speech. In Section 2

we indicated prosody as a possible source for cues of attribution. This section details how prosodic information can be used to identify ARs in speech.

Example 2 presents an utterance transcribed from SARC where ARs could be present. Considering only the lexical level, however, the sentence could be subject to many possible interpretations (e.g. there are at least 3 different possible lexical cues (represented by verbs of saying)).

(2) I *said* to him when you left do you remember I *told* you I *said* to him don't forget Dave if you ever get in trouble give us a call you never know your luck.

To choose the correct interpretation, we employed the judgement of a human annotator who listened to the recording and then we conducted an analysis of the acoustic features suggested in previous studies using Praat (Boersma and Weenink, 2014).

As shown in Figure 2, the waveform of the reported example is divided into two phases by a pause (0.7 seconds) (between the dotted lines) which occurs between the second *I said to him* and *don't forget*. The presence of the pause seems to mark the beginning of a new prosodic unit, which, directly following the lexical cue *said*, could be reported speech. The two graphs in Figure 3

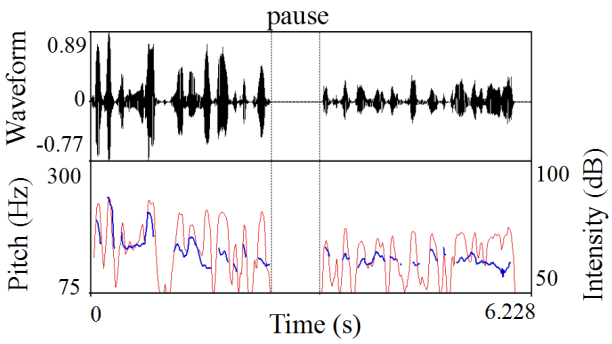

Figure 2: Rawdata of Ex(2).

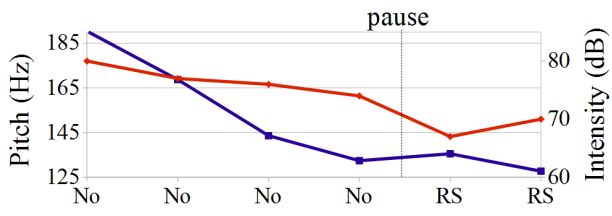

Figure 3: Means of Pitch and Intensity in Ex(2).

shows the variation in the means of respectively pitch (Hz)(blue) and intensity (dB)(red) along the timespan of the excerpt, elaborated from the rawdata in Figure 2. On the x-axis is displayed the presence (RS) or not (No) of reported speech according to our intepretation of the pause (dotted line) marking. The means of pitch and intensity show a similar tendency: a decrease of the mean with a stabilisation to a lower level after the pause. All the acoustic features seem therefore to suggest a difference in the prosodic marking between the first time span (No) and the second one (RS). This interpretation matches the one given by the human annotator. Thanks to the integration of the lexical cues with the acoustic analysis of the three prosodic factors combined it was possible to achieve the correct identification of the quotation (*don't forget Dave if you ever get in trouble give us a call you never know your luck*) out of at least three possible interpretations (considering only the verbs of saying). The full corpus contains many similar examples which demonstrate the importance of accessing to the acoustics for disambiguation of ARs in speech and how the judgements of human annotators can be analyzed by looking at the prosodic features.

6 Conclusions and future work

The analysis of SARC, the first resource developed to study ARs in speech, has helped to highlight a major problem of detecting attribution in a spoken corpus: the decreased reliability of the lexical cues crucial in previous approaches (completely useless in at least 10% of the cases) and the consequential need to find reliable prosodic clues to integrate them. The example provided in Section 5 has showed how the integration of the acoustic cues could be useful to improve the accuracy of attribution detection in speech.

As a future project we are going to perform a large acoustic analysis of the ARs found in SARC, in order to see if some reliable prosodic cues can in fact be found and used in order to develop a software able to extract attribution from speech.

Acknowledgments

We wish to thank the NLP group at the School of Informatics of the University of Edinburgh, where this project has been developed thanks to the Erasmus Placement Scholarship, and especially Bonnie Webber, Bert Remijsen and Catherine Lai.

References

Mikhail M. Bakhtin. 1981. *The dialogic imagination: Four essays*. University of Texas Press.

Claude Barras and Edouard Geoffrois and Zhibiao Wu and Mark Liberman. 1998. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. *First International Conference on Language Resources and Evaluation (LREC)*. 1373–1376.

Sabine Bergler, Monia Doandes, Christine Gerard and René Witte. 2004. Attributions. *Proceedings of the Eight International Conference on Language Resources and Exploring Attitude and Affect in Text: Theories and Applications, Technical Report SS-04-07*. 16–19.

Roxane Bertrand, Robert Espesser and others. 2002. Voice diversity in conversation: a case study. *Proceedings of the 1st International Conference on Speech Prosody*. Aix-en-Provence, France.

Paul Boersma and David Weenink. 2014. Praat: Doing Phonetics by Computer [Computer Program]. Version 5.3.63. Available online at: http://www.praat.org/.

Galina Bolden. 2004. The quote and beyond: defining boundaries of reported speech in conversational Russian. *Journal of pragmatics*. Elsevier. 36(6): 1071–1118.

Alessandra Cervone. 2014. Attribution Relations Extraction in Speech: A Lexical-Prosodic Approach. Thesis of the Master in Theoretical and Applied Linguistics. University of Pavia, Pavia.

David K. Elson and Kathleen McKeown. 2010. Automatic Attribution of Quoted Speech in Literary Narrative. *AAAI*.

Edouard Geoffrois and Claude Barras and Steve Bird and Zhibiao Wu. 2000. Transcribing with Annotation Graphs. *Second International Conference on Language Resources and Evaluation (LREC)*. 1517–1521.

Wouter Jansen, Michelle L Gregory, Jason M Brenier. 2001. Prosodic correlates of directly reported speech: Evidence from conversational speech. *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*. Molly Pitcher Inn, Red Bank, NJ, USA.

Gabriele Klewitz and Elizabeth Couper-Kuhlen. 1999. *Quote-unquote? The role of prosody in the contextualization of reported speech sequences*. Universität Konstanz, Philosophische Fakultät, Fachgruppe Sprachwissenschaft.

Ralf Krestel. 2007. Automatic analysis and reasoning on reported speech in newspaper articles. Tesis de Magister Universität Karlsruhe. Karlsruhe. Available at http://www. semanticsoftware. info/system/files/believer. pdf.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. *Proceedings of the Eight International Conference on Language Resources and Corpus Technology and Language Pedagogy: NewResources, New Tools, New Methods*. 3: 197–214.

Miguel Oliveira, Jr. and Dòris A. C. Cunha. 2004. Prosody as marker of direct reported speech boundary. *Speech Prosody 2004, International Conference*.

Silvia Pareti and Irina Prodanof. 2010. Annotating Attribution Relations: Towards an Italian Discourse Treebank. *Proceedings of LREC10*.

Silvia Pareti. 2012. A Database of Attribution Relations. *Proceedings of the Eight International Conference on Language Resources and Evaluation*. 3213–3217.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irene Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. *Proceedings of the 2013 Conference in Empirical Methods in Natural Language Processing*. 989–999.

Silvia Pareti. Forthcoming. Attribution: A Computational Approach. PhD Thesis, School of Informatics, the University of Edinburgh.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi and Bonnie Webber. 2007. Attribution and its annotation in the Penn Discourse TreeBank. *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse*. Citeseer. 47(2): 43–64.

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

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