"Why do you ignore me?" - Proof that not all direct speech is bad

Laura Hasler

Research Group in Computational Linguistics
School of Humanities, Languages and Social Sciences
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
L.Hasler@wlv.ac.uk

Abstract

In the automatic summarisation of written texts, direct speech is usually deemed unsuitable for inclusion in important sentences. This is due to the fact that humans do not usually include such quotations when they create summaries. In this paper, we argue that despite generally negative attitudes, direct speech can be useful for summarisation and ignoring it can result in the omission of important and relevant information. We present an analysis of a corpus of annotated newswire texts in which a substantial amount of speech is marked by different annotators, and describe when and why direct speech can be included in summaries. In an attempt to make direct speech more appropriate for summaries, we also describe rules currently being developed to transform it into a more summary-acceptable format.

1. Introduction

In the automatic summarisation of written texts, direct speech is largely ignored or deemed unsuitable when it comes to selecting important/appropriate sentences.[1] There are several reasons for this treatment of speech.

Humans do not usually include direct speech quotations when writing summaries, and the aim of many systems is to produce a human-like condensed text.[2] In linguistic analyses of texts (preliminary work; (Goldstein et al., 1999)), quotations are often used to elaborate statements already presented, and so are seen as redundant information taking up valuable space. If they do provide new and important information, they often contain opinions, together with personal pronouns (*I, we, you*) which may not be resolved within the summary sentences. Because direct speech is separated from the rest of the text by a specific type of punctuation, and is distinguished from other (the author's) text by the reader, it is not as easy or quick to read and process. As summaries are meant to be time-saving devices, this needs to be taken into consideration.

However, it is possible that not all direct speech is bad. There are occasions where it does contain new, important information, without too many (or any) opinions or unresolvable references and can therefore be useful to include in a summary. Through an examination of a corpus of newswire texts annotated for important information, this paper argues that it is not always safe to discard direct speech *per se*, and that potentially useful information can be excluded from summaries by ignoring such text.

The paper is structured as follows. Section 2 details related work and the general attitude to direct speech in written summaries. The corpus is briefly described in section 3, followed by a discussion of the results of the corpus analysis in section 4. Some transformation rules for speech sentences are presented in section 5, and the paper finishes with conclusions and ideas for future work.

---

[1]The type of summarisation referred to in this paper is the creation of extracts, not abstracts.

[2]For more information on the ways in which humans summarise texts, see (Endres-Niggemeyer et al., 1995).

2. Direct speech in summaries of written texts

Speech is often ignored during the formulation of rules and annotation guidelines based on linguistic analyses of texts. This means that systems or methods which use guidelines for human annotation of important sentences when building corpora for summarisation, or that use rules based on these corpora to determine the text to be selected, exclude what can actually be important information.

(Goldstein et al., 1999) discuss the fact that in their data (articles and summaries from Reuters and the Los Angeles Times), 'Words and phrases common in direct or indirect quotations tended to appear much more frequently in the non-summary sentences'. As these are empirical observations from human-written summaries, they seem a good starting point for annotation guidelines for summarisation, especially if the data comes from a similar source. However, this is not borne out by an analysis of the corpus presented in (Hasler et al., 2003), as direct speech is selected as important by all annotators.

Although their guidelines do follow others in some respects, taking a similar attitude to direct speech and stating that it should not be marked as important (unless it presents new and vital information), a substantial number of the texts examined contains at least one case of direct speech marked as important. All 4 annotators marked direct speech as important on at least two occasions. Given the guidelines, those units that are marked can be considered particularly relevant because they had to compete with other (preferred) non-speech text. This suggests that we cannot ignore direct speech, at least in the summarisation of written newswire texts, and that this needs accounting for so that in future, important information is not omitted purely because it appears inside quotation marks. The fact that the instances of marked speech in the corpus are not limited to one or two provides more evidence for this.

3. The corpus

The corpus used in this investigation is taken from (Hasler et al., 2003), which contains newswire documents

from the Reuters corpus (Rose et al., 2002) annotated for summarisation by 4 different annotators. Of the 147 newswire texts in this corpus, 44 contain at least one instance of direct speech marked as important. These 44 texts (1282 sentences) are the basis for the analysis here. They comprise business, politics and sports texts, and are taken from all 4 annotators. Table 1 shows the basic statistics of our direct speech documents corpus. The corpus is described in numbers of sentences rather than numbers of words, as this is the unit specified for annotation, and the unit used in the majority of summarisation systems.

We consider marked and unmarked speech from the same document, the classification of the marked speech (essential/important/referred), reporting clauses, pronouns and references in marked and unmarked direct speech sentences, and differences between the sub-genres in the corpus. An explanation and discussion of these can be found in section 4.

	Business	Politics	Sport	Total
Texts	27	12	5	44
Sentences	752	342	188	1282
Speech sentences	169	84	103	356
Marked speech	31	21	27	79
+ r-clause	10	17	7	34
- r-clause	15	4	4	23
no r-clause	0	0	9	9
previous r-clause	6	0	7	13
Unmarked speech	138	63	76	277
Essential	8	6	12	26
Important	17	11	14	42
Referred	6	4	1	11

Table 1: Direct speech texts corpus statistics: sentences

4. Corpus Analysis: Results and Discussion

4.1. Functions of direct speech

There are three basic, underlying functions of direct speech in the corpus. The first is where the fact that the speech is spoken by a certain person is important, for example, to give opinions on information presented in the text. Secondly, there is speech which functions as a standard non-speech sentence, i.e. as a statement of fact which, when its quotation marks and reporting clause (if present) are removed, does not necessarily need to be attributed to any speaker. The third type of speech elaborates and provides supporting quotations to emphasise or support information already presented. It also tends to include opinions, but ones which are less relevant. Within these broad categories, there are also more specific roles fulfilled by speech. However, when marked as suitable for inclusion in a summary, direct speech has a different basic function to speech which is not considered thus.

All marked sentences should provide important information about the main topic(s) of the text, as specified in the annotation guidelines.[3] In most cases, this is true

---

[3]The term *topic* is used in its most general sense in this paper.

(see section 4.1.1 for other cases). Marked direct speech is either speech functioning as non-speech, or speech providing opinions where opinions and their speaker are important to the overall understanding of the main topic. In most cases, marked speech in a text supplied one or two opinions, in single sentences. However, sports texts contain a higher proportion of speech sentences, which contain features more typical of speech than those in the other sub-genres, including sequences of speech sentences which can be more repetitive than other text. This means that, although it is not true in every case and there is an element of sub-genre dependence, the function that speech assigns itself can help determine the feasibility of its inclusion in a summary. Speech mostly functioning as the third type of speech mentioned above is not marked as important.

4.1.1. Classification of marked speech

According to (Hasler et al., 2003)'s annotation guidelines, those sentences worthy of inclusion in a summary are classified as essential, important or referred. Referred sentences are those which are not important themselves, but which need to be included to ensure a full understanding of other marked sentences. They may contain, for example, a noun phrase or the explanation of an acronym later referred to in the text.

From the figures in Table 1, we can see that the main reason for marking speech is that it actually contains important information which needs to be included in the summary. As the annotators preferred to select these speech sentences over alternative non-speech sentences, then the important information in them must be the most appropriate information, expressed in the best way, because the guidelines advise not to select speech where possible.

Although there are some marked speech sentences classified as referred, and introducing items which are later referred to in other marked sentences is obviously one thing that speech does, it was not the main reason for marking speech for inclusion. However, when sentences are marked as referred, it is not usually the content of the speech sentence which is the reason for its annotation. In most cases it is the speaker of the sentence which is important because they are referred to later.

4.2. Characteristics of marked speech

As mentioned above, speech is usually marked as important because it is precisely that. The annotators' comments in a number of files that "this is speech but it is more important than the text which isn't speech" emphasise this point. The information contained in marked speech sentences relates to the main topic of the text, and either is not present in non-speech sentences or is presented more appropriately, i.e. it is the most succinct version of relevant information, in the speech sentence. Much of the marked speech provides a good one sentence summary of the main meaning or sentiment behind the text.

Pronouns are present in marked speech sentences just as much as in unmarked sentences, with one main difference: all the pronouns within marked speech sentences are resolvable within the other important sentences. This means that no additional sentences, which increase both

summary length and amount of irrelevant and repetitive material, need to be taken into account (apart from those already marked as referred, where necessary).

Opinions are sometimes included, mainly in the form of *I think*, and especially in sports texts. There is more opinion in these texts than the others, as well as more speech (see section 4.4), which means that both of these are more likely to be included in important sentences within these texts. The importance of opinion depends very much on the individual text, for example it is more important where there are two conflicting sides. The information in the sentence can also be important regardless of the fact that an opinion is present. Speech also works to clarify, for example, if the text is littered with speculation or confusion, a direct quotation (including the reporting clause) can succinctly summarise an official view or the viewpoint of the main protagonist, or clear up the surrounding confusion.

4.3. Reporting verbs and speakers

4.3.1. Presence of reporting clause

Speech sentences which keep their reporting verb and speaker are most frequent (see Table 1). As the annotators had the option to remove segments which carried no important information from sentences, we can assume that those reporting clauses which are kept were deemed important in themselves by the annotators.

The reporting clause is kept for two main reasons. Firstly, the speaker themselves is important; the reader needs to know who exactly to attribute the speech to for the text to make sense, there may be a personal pronoun such as *I* or *we* present, or the speaker adds justification and credibility to a point being made, for example, if the information comes from an official source. This is most obvious in documents where the main point of the text is to convey different views about a particular event, or if the text contains lots of speculation or opinion. Secondly, the reporting clause contains important information about the speaker, a place or a situation which is not present elsewhere. We should remember that this information is considered important in addition to the content of the speech sentence, and is not the only reason for marking the sentence. To illustrate:

> "There was no battle for the city and no looting," said one source in Lubumbashi, the capital of Shaba province in the far south-east copperbelt and one of Kabila's next declared targets.

Speech sentences classified as referred due to their speaker are another case of when the reporting clause is kept (see section 4.1.1).

4.3.2. No reporting clause

There are three different types of speech sentences without a reporting clause: i) the reporting clause is removed by the annotator, ii) the reporting clause can be inferred from the previous speech sentence, iii) there is no reporting clause and it cannot be inferred. The most interesting type to us here is type i), and what we are concerned with is why the annotator removed this clause. Most of this type of speech is the same as a non-speech

sentence when the reporting clause and quotation marks are removed; it functions as a simple statement of fact and there are no "typical speech features" (such as opinions or lots of personal pronouns) that could make it difficult to understand without its reporting clause. When this is the case, it is not important that the reader knows who the actual speaker is. Another reason for the removal of this clause is that its information is already included in a previous marked sentence and to keep it in the second (or third) instance would make the summary repetitive.

4.4. Sub-genres

Within our corpus, there are texts from three different sub-genres: business, politics and sport. The sub-genre of a text affects most of the things we have investigated here.

Sports texts contain by far the most speech, and hence the most marked speech sentences. In these texts, speech is such a widespread feature that it has to be selected in order to obtain the most important information. All the speech sentences except one are considered important or essential and are not included because their speaker is referred to later. There is an even distribution of essential and important classification. There is also a more even distribution of the keeping and removal of reporting clauses, and more sentences with no speaker or speaker inferrable which is due to more sequences of speech sentences from coaches and players.

Politics and business texts contain a much smaller but similar amount of marked speech in terms of the total number of sentences (speech and non-speech), although politics texts have a higher percentage of speech sentences that are marked. Treatment of the reporting clause differs, being kept much more often in the sub-genre of politics where the individual speaker is important to the meaning of the text, than in the more descriptive, one-sided business texts which do not need to include such information. In these sub-genres, more marked speech sentences are classified as important than as essential or referred, and there are more classified as referred than in sports texts.

4.5. Unmarked speech

Unmarked speech generally supports information which has already been presented or information secondary to the main topic of the text. If this is information in a marked sentence, then it provides more credibility and additional evidence from a reliable source to help justify the story. This speech is unmarked because it repeats information already presented, or elaborates marked sentences, providing too much additional detail, or presents obvious conclusions which can easily be drawn from the information already marked.

Unmarked speech also contains references, mainly to people, (NPs, pronouns such as *he*, *we*, *I*) which cannot be resolved within the existing marked sentences. Other sentences would also need to be marked in order to resolve these references, making the summary much longer, as well as adding a substantial amount of irrelevant information which constitutes the main part of the additional sentences. A clear example of this is also linked to the amount of opinion in unmarked speech sentences. There are a number

of occurrences of *I think* and *we think* attributed to more than one speaker, which as well as being difficult to resolve, are irrelevant because they tend to give more personal views than are generally relevant in summaries.

There is a strong association of future speculation and speech, as well as details of emotions and feelings (*I feel good*), and reactions, which are not usually central to the main topic of the text. There is also an element of negativity in the unmarked speech sentences, with a higher number of phrases such as *I don't think*. Where opinion is deemed important (there are several cases, mainly in the sports texts), positive opinion is preferred to negative opinion. This does however, depend on the sub-genre, and to a certain extent the individual text itself.

5. Rules for transformation of speech

As ease and speed of reading is an issue in summarisation, and as direct speech may not be as easy or quick for the reader to process due to its dissimilarity with the surrounding text, it is preferable for direct speech from the source text to be represented in a summary as indirect speech or as a standard declarative sentence, as appropriate. The rules currently being developed aim to address this and although space restrictions prevent the explicit justification of each rule here, they are based on the corpus analysis of marked speech sentences described above (section 4). Additional rules are also being formulated (see section 6 for future plans); our preliminary rules to date are presented below.

All marked speech sentences: punctuation
Remove " "
Replace **,** at end of speech with **.** or add **.** if nothing present
Remove **,** at beginning of speech

All marked speech sentences: pronouns
I $\longrightarrow$ *he/she/antecedent*, change verb accordingly: "As far as I know both sides have retired... " he said $\longrightarrow$ *He said as far as he knows both sides have retired...*
my $\longrightarrow$ *his/her*
our $\longrightarrow$ *their*
us $\longrightarrow$ *them*
ourselves $\longrightarrow$ *themselves*
we (except generic, inclusive *we*) $\longrightarrow$ *they/antecedent*

Reporting clause kept
X said (and variants)
If reporting clause is placed after speech, reposition before speech: ... " a dealer said $\longrightarrow$ *a dealer said...*
said X (and variants)
Swap verb and speaker position: ... " said Laluci $\longrightarrow$ *Laluci said...*
That
Add *that* after *added*: ... " he added $\longrightarrow$ *he added that...*
Add *that* if other text is placed between reporting verb and speech: ... " Prime Minister Romano Prodi told reporters $\longrightarrow$ *Prime Minister Romano Prodi told reporters that...*
... " Eirvin Knox, said in an interview on Tuesday $\longrightarrow$ *Eirvin Knox said in an interview on Tuesday that...*

6. Conclusions and future work

This paper is concerned with a corpus analysis of direct speech sentences considered by human annotators to be suitable summary material. Although direct speech is usually ignored in the summarisation of written texts, through this investigation of marked and unmarked speech, we have identified circumstances when speech is suitable for inclusion in a summary and when it keeps/loses its reporting clause. We have proved that not all direct speech is bad. The speech annotated in this corpus contains important information about the main topic(s) of the texts, without too many instances of irrelevant background or secondary information, or too many features typical of speech such as emotion, repetition and opinion. If this speech is not taken into account, some of the most important information in the texts will be missed. Therefore, future annotation guidelines and rules for the selection of important information should not discriminate against these types of direct speech sentences.

We have also developed a preliminary set of rules to transform speech sentences into non-speech sentences, which better suit the general summary style. These rules are being developed further to deal with other aspects of speech. When *we* is used generically or inclusively, it does not need to be transformed in the same way as other pronouns, and so we need to be able to reliably distinguish between the different uses. There are also some typical speech-style phrases in the marked speech sentences, such as *you know* that does not carry any real meaning; it is important to differentiate between these phrases and phrases containing the same words but carrying a different meaning. Other interesting aspects we are considering are the use of direct quotations in indirect speech sentences, and how to deal with the sometimes random repetitions that occur in direct speech. Investigation into these is already underway.

7. Acknowledgement

8. References

Endres-Niggemeyer, B., E. Maier, and A. Sigel, 1995. How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor. *Information Processing and Management*, 31(5):631 – 674.

Goldstein, Jade, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Proceedings of SIGIR '99*. Berkeley, California, USA.

Hasler, Laura, Constantin Orăsan, and Ruslan Mitkov, 2003. Building better corpora for summarisation. In *Proceedings of Corpus Linguistics 2003*. Lancaster, UK.

Rose, T. G., M. Stevenson, and M. Whitehead, 2002. The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002)*. Las Palmas de Gran Canaria.