Extraction of Unmarked Quotations in Newspapers

A Study Based on Direct Speech Extraction Systems

Stéphanie Weiser, Patrick Watrin

CENTAL - Institut Langage et Communication - UCLouvain
1348 Louvain-la-Neuve, Belgium
{stephanie.weiser, patrick.watrin}@uclouvain.be

Abstract

This paper presents work in progress to automatically extract quotation sentences from newspaper articles. The focus is the extraction and annotation of unmarked quotation sentences. A linguistic study shows that unmarked quotation sentences can be formalised into 16 patterns that can be used to develop an extraction grammar. The question of unmarked quotation boundaries identification is also raised as they are often ambiguous. An annotation scheme allowing to describe all the elements that can take place in a quotation sentence is defined. This paper presents the creation of two resources necessary to our system. A dictionary of verbs introducing quotations has been automatically built using a grammar of marked quotations sentences to identify the verbs able to introduce quotations. A grammar formalising the patterns of unmarked quotation sentences – using the tool Unitex, based on finite state machines – has been developed. A short experiment has been performed on two patterns and shows some promising results.

Keywords: Unmarked Quotation Sentences, Information Extraction, Annotation

1. Introduction

Quotations are very common in newspaper articles and their identification can be useful for many applications such as automatic summaries, biography creation or opinion mining. If they are often signaled by quotations marks, some are not and some appear in indirect reported speech. In the context of the Biographe project[1], whose goal is to automatically build biographies of famous people using information found on the Web (Wikipedia and newspaper articles), the extraction of quotations is essential: biographies can then be completed by "what the person said" and by "what was said about the person".

In this paper, we present work in progress which targets the automatic extraction and annotation of all types of quotations (direct / indirect, marked / unmarked) from newspaper articles in French. This paper focuses on the extraction of unmarked quotations: creation of a grammar of unmarked quotations and automatic construction of a dictionary of verbs that can introduce quotations.

Section 2 briefly presents the state of the art in extraction of quotation sentences and shows what we mean by the term "quotation". In section 3, a study of unmarked quotations is performed and their syntactic structures are formalised. The linguistic difficulties of insertions and boundaries are described. The annotation scheme we developed for quotation sentences annotation is then described. Section 4 presents the resources we built to automatically extract the unmarked quotations: grammar of quotation sentences and dictionary of quotation verbs. Section 5 provides the preliminary figures that our experiment yielded.

[1]Project for the creation of biographies in Dutch, English, French and German developed at the Cental.

2. State of the Art

Many systems that automatically extract quotations from newspaper articles already exist for English, French, and other languages, like News Explorer (Pouliquen et al., 2007) or Excom[2] (Alrahabi and Desclés, 2008). However, they mainly concern direct speech quotations, signalled by quotation marks. If some researchers have also worked on indirect speech or unmarked direct speech recognition (Danlos et al., 2010; Sagot and Danlos, 2010), no thorough study has been done that provides an evaluation corpus and comparison data. Reported speech has been widely studied in linguistics (Maingueneau, 1996; Charaudau and Mainguenau, 2002) and we intend to base our research on strong linguistic foundations. However, as it will be shown in the next section, we adopted a wider definition of "quotation", in order to meet our applicative needs. We distinguish between marked and unmarked quotations (depending on the use of quotation marks). Following the classic definitions, marked quotations often correspond to direct reported speech, although unmarked quotations can present the same structure. This is why we prefer to use the marked / unmarked distinction.

3. Linguistic Study

This section describes shortly the corpus we have used and presents the linguistic study we have conducted: formalisation of the structures of unmarked quotation sentences and identification of the boundaries of the reported part. The annotation scheme we have defined is then introduced.

3.1. Corpus

A newspaper corpus has been chosen for this study for two reasons. First, the type of texts found in newspapers is very

[2]http://www.excom.fr

likely to contain quotations. Second, it is the detection of quotations, precisely in newspapers, that could lead to the development of a useful tool. The corpus used for this study contains around 95 000 press dispatches collected in 2003 from the Belga press agency.

3.2. Unmarked Quotations Structures

In newspaper articles, quotations are introduced in many different ways. Our work focuses on the unmarked quotations that are introduced by speech verbs (such as *declare, say, announce...*)[3]. Sentences like the two following are typical of the kind of expressions that need to be extracted:

> Le bilan définitif des morts en Thaïlande dus au raz-de-marée de dimanche pourrait être de 7.000 à 8.000 personnes, a annoncé samedi le Premier ministre Thaksin Shinawatra.
>
> [*The final death toll in Thailand due to the tidal wave on Sunday could be between 7,000 and 8,000 people, announced the Prime Minister Thaksin Shinawatra, on Saturday.*]
>
> Plus de 1.000 Norvégiens pourraient avoir été tués, a déclaré le Premier ministre norvégien Kjell Magne Bondevik.
>
> [*More than 1.000 Norwegians may have been killed, stated the Norwegian Prime Minister Kjell Magne Bondevik.*]

The output of the process presented here is an unlexicalised grammar which formalises the syntactic structures of unmarked quotations. Section 4 explains how we semi-automatically lexicalise this grammar to produce the final grammar.
This section will show that the two main linguistic difficulties are the definition of the structures, with the detection of possible insertions, and the identification of the boundaries of the quotations.

3.2.1. Formalisation (Structures Definition)

In carrying out a linguistic study of unmarked quotations, we chose three verbs that very frequently introduce quotations and that are very productive: "déclarer" (*to declare*), "annoncer" (*to announce*) and "ajouter" (*to add*). We manually studied the context in which these three verbs appear in our corpus and selected examples where they are used to introduce unmarked reported speech. These examples were analysed in order to formalise unmarked reported speech and build structures that could be generalised. We defined 16 structures of unmarked reported speech sentences. These structures are based on the elements that take place in the quotation sentence and their order: the verb introducing the quoted text, the quotation itself, the reference to the person who is quoted and other elements that can be inserted in the same sentence, like a temporal or spatial information. The following two structures are quite simple and very frequent. They contain the three elements that a sentence must contain to qualify as a quotation sentence.

1. [Unmarked Reported Speech], [Verb] [Person Name].
2. [Person Name] [Verb] [Unmarked Reported Speech].

The unmarked reported speech parts can contain any words except the verbs "déclarer", "annoncer" and "ajouter", and the punctuation marks ", ; :".
Some insertions can appear in the sentence (between the verb and the person's name, at the end or the beginning), such as time information (day of the week), location, the addressee's name, adverbs (e.g. *solemnly*)... In order to find the boundaries of the reported speech, these insertions need to be identified.
Some of the structures for unmarked reported speech are more complex, such as:

3. [Unmarked Reported Speech] [Verb] [Person Name], *ajoutant (adding)* [Unmarked Reported Speech].
4. [Person Name] [Verb] *que (that)* [Unmarked Reported Speech] - [Marked Reported Speech] *et que (and that)* [Unmarked Reported Speech] - [Marked Reported Speech].

The structure number 3 shows that the reported speech can be split into two parts. The structure shown in 4 shows that unmarked reported speech can alternate with marked reported speech.

3.2.2. Boundaries of the reported part

The main difficulty encountered with unmarked reported speech identification is to find the boundaries of the reported part, a problem that obviously does not concern marked reported speech. However, this difficulty is applied to indirect reported speech in general, not only for automatic processing. For example, in the following sentences, it is not possible for a human reader to determine for certain the end of the reported part.

> Il a ajouté que le ministère public a défendu la même position devant la chambre des mises en accusation, qui se réunit à huis clos.
>
> [*He added that the public prosecutor defended the same position before the court's indictment division, which meets in camera.*]
>
> Il a ajouté par ailleurs que cinq camions transportant du matériel de secours étaient arrivés dans la journée à Bam depuis la ville de Kermanshah, située à plusieurs centaines de km de distance.
>
> [*He also added that five trucks transporting rescue equipment had arrived in Bam during the day from the town of Kermanshah, situated several hundred kilometres away.*]

In the two sentences, the last part *qui se réunit à huis clos (which meets in camera) / située à plusieurs centaines de km de distance (situated several hundred kilometres away)* can either be part of the quotation itself or be a comment of the journalist, author who reports this speech.
With our manual study of many unmarked quotations, we have noticed that, in most of the cases, when the boundaries of the reported speech are not ambiguous, it goes as far as

[3] Quotations can also be introduced by expressions such as *according to someone*. These structures seem to form a finite set of expressions that can be formalised in a grammar. Here we mainly focus on quotations introduced by verbs

the end of the sentence. We have assumed that this rule can be generalised to the cases where they are ambiguous and it is this rule that we have implemented in our grammar.

3.3. Annotation Scheme

The description of the structures is enough to build an extraction grammar (as the one presented in section 4.1.). However, a complete annotation scheme to define the annotation format is fundamental in order for the identified quotations to be used in a larger application. Moreover this annotation scheme can be used to annotate a corpus and create a gold standard.
We chose the XML format to annotate the quotations and defined a DTD to describe the annotation scheme. The elements of this scheme are the following:

- <AUTHOR> – to mark who is quoted: it can be a person name, a pronoun, a collective element (*x, y and z have said...*), an abstract element (*the journal, the report, the police...*)
- <QUOTE> – the quoted part
- <VERB> – the verb introducing the quotation
- <INTRO> – to mark the element introducing the quotation when it is not a conjugated verb (*according to...*)
- <INSERTION> – to mark all the other information that can appear in the sentence (like temporal or spatial information...)

The two following tags allow to regroup elements:

- <RS> (reported speech) which aim is to contain all the parts of the quotation sentence.
- <QV> which links the verb to the quotation.

The tag <RS> is used to mark the whole sentence, it can therefore contain more than one <QV> when more than one reported part is included in the sentence. The only, but very important condition is that all the quotations included in one <RS> need to have the same author.

4. Resources

The two resources that are needed to automatically identify unmarked quotations are an implementation of the formal grammar of quotation structures described in Section 3 and a dictionary of verbs able to introduce reported speech.

4.1. Creation of a Grammar

The grammar that we need to build for this work has to recognise the 16 syntactic structures that we have defined for unmarked quotations. It is unlexicalised: it describes only the structures of quotation sentences but does not contain lexical data. For our first experiment, we built a grammar to extract unmarked quotations following the structures 1 and 2 presented in 3.2.1. We used the Unitex[4] tool, base on finite state automata, to build the grammar. This grammar should be extended to match the 16 structures but it already shows our methodology yields positive results. It should also be extended to identify the different types of insertions.
These structures are (too) generic and using them as they are would introduce a lot of noise. The grammar then needs to be lexicalised and this task concerns mainly the verbs that introduce quotations. A list of the most frequent verbs could be created manually with a short corpus study, but this would not include the verbs that are less frequent but nevertheless substantially represented, like *susurrer* (∼*to whisper*) or *prophétiser* (∼*to prophecy*). This is why we used the following method to build our dictionary.

4.2. Dictionary of Quotation Verbs

We suspected that the verbs used to introduce unmarked quotations could be the same ones that introduce marked quotations. As a naive approach can be followed to extract, quite efficiently, marked quotations, we have applied it to build a corpus of sentences containing quotations. We identified all the sentences containing a pair of quotation marks and a verb (outside of the quotation marks). Sentences like the following where therefore identified:

> "I think Rick Perry's boomlet probably really peaked in August and has subsided," Jackson said.

Each sentence was annotated, and then all full verbs situated within a 5-word window to the left and right of the quotation were extracted. The list of extracted verbs can directly constitute a dictionary. However, some analysis and work on this list could help to improve it, for example in erasing the verbs that are not frequent enough to be significant. Moreover, the "frequency" criterion is not sufficient and an association measure should be more efficient: we will use a semi-automatic method to determine the verbs that are used most frequently to introduce quotations rather than for other purposes.

5. Experiment

The experiment we performed was very short. However, it shows that our methodology has the potential to obtain good results, once we build a more complete quotation extractor.
We applied our grammar for the first two structures to our press dispatch corpus, using the three chosen verbes (*déclarer, annoncer, ajouter*). The first limitation of this experiment is that, since people's names were not annotated beforehand, we could only match the cases where the person's name contained a first name and only one other word. However, it has the advantage of matching only 38 expressions for the first structure and 102 for the second, which allowed us to count them and analyse them thoroughly.

5.1. A Few Numbers

These numbers do not give the opportunity to significantly calculate precision and recall measures. However, they still show the quality of the structures.
As is shown in Table 1, for the first structure, on 38 matches, 30 quotations were correctly identified (with correct boundaries). 8 expressions were considered as quotations when they were not or were only a partial match on a

[4] http://igm.univ-mlv.fr/ unitex/

	Structure 1	Structure 2
Matches	38	102
Correct	30 (78,9%)	76 (74,5%)
Noise	8 (21,8%)	8 (7,85%)
Unclear boundaries	N/A	18 (17,65%)

Table 1: Preliminary figures for unmarked quotation extraction

quotation. For the second structure, 102 expressions were identified as quotations. It was correct for 76; 8 expressions should not have been identified; and for 18 expressions, the boundaries of the quoted part do not seem clear (for a human reader).

5.2. Results Analysis

The expressions that were wrongly considered as quotations with the first structure were composed of only a few words:

> Cette vidéo / Mais (2x) / Néanmoins / Pour nous / Qui plus est / Enfin / Par ailleurs
>
> *This video / But / however / For us / Moreover / Finally / In addition*

Only the first one ("Cette vidéo" (*This video*)) is, for sure, part of a quotation:

> Cette vidéo, a ajouté Mike Currie du centre spatial Johnson à Houston (Texas) "commence après la mise à feu [...]
>
> *This video, added Mike Currie of the Johnson spatial center in Houston (Texas) "begins after the firing [...]*

Rules to analyse the content of the identified quoted part could help to eliminate this type of error: for example, adverbs, before an introducing verb should not be considered as a complete quotation. It could however be included in the following quotation. It is described in one of the other structure patterns and its implementation into the grammar would probably avoid some of these errors.

For the second structure, the results are more varied. Some of these sentences do contain unmarked quotations but follow another structure and the selected part is therefore not the right one. In one case, the verb "ajouter" (*to add*) was used in a mathematical sense and not as an introducer of reported speech. In the other cases, the verb "déclarer" was used in a larger expression "déclarer forfait" (*to forfeit*) which also does not introduce reported speech. Some of these mistakes could therefore easily be avoided by improving the grammar, including elements to exclude.

6. Conclusion and Future Work

Of course, in order to build a fully operable system, we need to complete the grammar and improve the dictionary. Our intention is to use dictionaries as much as possible in order to separate the grammar describing the structures of the information to extract from the linguistic data. By doing so, we will be able to test our system on other languages than French, by creating the dictionaries, without modifying the grammar too much.

We have now applied the grammar and dictionaries developed for the Biographe project on a corpus of 400 news articles. In the future, this step will be prior to the quotation annotation process and the grammar developed will rely on the pre-annotated entities. Many entities are annotated in the corpus: person names, professions, temporal informations, organisations, places... Based on this pre-annotated corpus, we are currently creating a gold standard corpus annotated manually, following the annotation scheme described in section 3.3. For the annotation process to be more effective, the tool Glozz[5] (Widlöcher and Mathet, 2009) will be used. The annotation scheme therefore needs to be adapted to the Glozz format and stand-off annotations will be provided by the tool but these formats are compatible with the XML technologies. The corpus created will be used as an evaluation corpus and it will help afterwards to improve the grammars.

A system able to annotate quotation sentences in news articles can be used in larger applications. For example it could be used to build a network of people based on the quotations, in which the people who are quoted a lot would be central. With some sentiment analysis techniques, this graph of people would also be oriented with positive and negative information to show *who likes who*.

7. Acknowledgements

We would like to thank the reviewers of the submitted abstract who have provided us with helpful advice.

8. References

Motasem Alrahabi and Jean-Pierre Desclés. 2008. Automatic annotation of direct reported speech in arabic and french, according to semantic map of enunciative modalities. In *6th International Conference on Natural Language Processing, GoTAL*, pages 41–51, Gothenburg, Sweden.

Patrick Charaudau and Dominique Mainguenau. 2002. *Dictionnaire d'analyse du discours*. Seuil. (DA), Paris.

Laurence Danlos, Benoît Sagot, and Rosa Stern. 2010. Analyse discursive des incises de citation. In *2ème Congrès Mondial de Linguistique Française - CMLF 2010*, page ., La Nouvelle Orléans États-Unis. Institut de Linguistique Française.

Dominique Maingueneau. 1996. *Les termes clés de l'analyse du discours*. Seuil. (DA), Paris.

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.

Benoît Sagot and Laurence Danlos. 2010. Verbes de citation et Tables du Lexique-Grammaire. In *International Conference on Lexis and Grammar*, Belgrade Serbie, 09.

Antoine Widlöcher and Yann Mathet. 2009. La plate-forme Glozz: environnement d'annotation et d'exploration de corpus. In *Actes de TALN 2009*, Senlis, France. A paraître.

[5] http://www.glozz.org/