2
Quotation Extraction

In this chapter, we present the Quotation Extraction task and the task decomposition. In addition, we show the task inherent difficulties which allow us to deduce that this is not a trivial task. We also present quotation types (34) and the types our work proposes to identify. Furthermore, we show related works and the peculiarities of each one. Finally, we present the differences between our work and the other ones.

2.1
The Task

The purpose of the Quotation Extraction task is to identify quotations in a text and associate them to their authors. We decompose this task into two subtasks: quotation identification and association between quotation and author.

> Nélio Machado que defende Daniel Dantas, considerou ‘*estranha*’ a acusação de que Dantas teria cogitado subornar o juiz. ‘*Isso é o fim da picada. Completamente sem fundamento e bem no dia em que o Daniel vai prestar depoimento. Estou inclinado a pedir suspeição dele [Fausto de Sanctis]. Acho muito estranho, tem conteúdo de mais armação do que qualquer outra coisa’* disse ele.

Figure 2.1: Quotation identification subtask

> **Nélio Machado**$_1$, que defende **Daniel Dantas**$_2$, considerou ‘estranha’ a acusação de que **Dantas**$_2$ teria cogitado subornar **o juiz**$_3$. ‘Isso é o fim da picada. Completamente sem fundamento e bem no dia em que o **Daniel**$_2$ vai prestar depoimento. Estou inclinado a pedir suspeição **dele**$_3$ [**Fausto de Sanctis**$_3$]. Acho muito estranho, tem conteúdo de mais armação do que qualquer outra coisa’ disse **ele**$_1$.

Figure 2.2: Author candidates for the quotation association to its author subtask.

In Figure 2.1, we show an illustrative example of the quotation identification subtask. Quotations are in *italic.*

For the second subtask, we provide the author candidates, which are the coreferences in the text. Figure 2.2 presents an example with three coreference sets. Each coreference is in **bold** and is tagged with an integer subscript which indicates its respective coreference set label.

> **Nélio Machado**$_1$, que defende **Daniel Dantas**$_2$, considerou '*estranha*'$_1$ a acusação de que **Dantas**$_2$ teria cogitado subornar **o juiz**$_3$. '*Isso é o fim da picada. Completamente sem fundamento e bem no dia em que o* ***Daniel***$_2$ *vai prestar depoimento. Estou inclinado a pedir suspeição* ***dele***$_3$ *[**Fausto de Sanctis**$_3$]. Acho muito estranho, tem conteúdo de mais armação do que qualquer outra coisa*'$_1$ disse **ele**$_1$.

Figure 2.3: Association between quotation and author subtask

In the association subtask, we associate the quotation to its respective coreference set label, which identifies the quotation author. In Figure 2.3, we show an example of two associations between quotation and author. Each quotation is tagged with its respective author coreference set label.

We present a diagram of the task decomposition in Figure 2.4. For the quotation identification subtask, we provide the input corpus with part-of-speech and named entity annotations. For the association between quotation and author subtask, we provide the input corpus with quotation and coreference annotations.

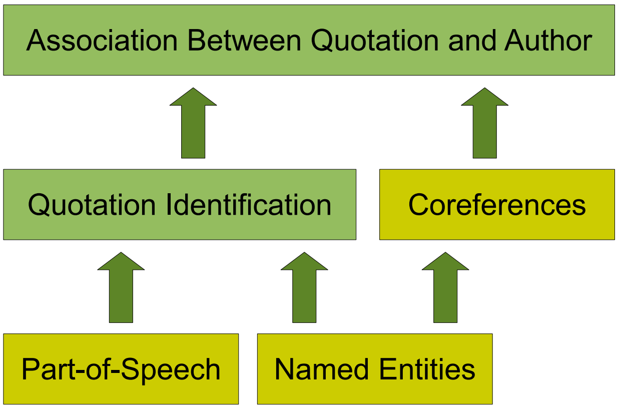

Figure 2.4: Task decomposition diagram

Named entity recognition (41, 36, 33, 35, 2) and coreference resolution (37, 8) are classical NLP tasks that do not have good quality. Thus, in order to prevent those tasks from impacting in our task negatively, we use a golden annotation for both named entities and coreferences as input to our Quotation

Extractor. For part-of-speech annotation, we use a state-of-the-art tagger (11) that possesses good quality.

1. *–Confio na minha absolvição. Não tem lógica ser indiciado por isso* – diz Schwenck.
2. “*Foi ótimo, tudo lindo*”, elogiou ela, que usava uma bolsa tipo carteira multicolorida da marca Prada.
3. Mantega se diz ‘*satisfeito*’ com comportamento da inflação
4. Segundo ele, esta foi apenas uma primeira fase. ‘*Agora, temos de identificar os funcionários públicos e privados que pediam os atestados porque isso prejudica alguém*’.

Figure 2.5: Several examples of quotations

1. Com uma atuação empolgante, e aproveitando-*se das falhas ofensivas da equipe do Los Angeles Lakers...*
2. O apelido ‘*Ilha do Retiro*’ foi dado em homenagem ao próprio bairro onde a arena fica situada.
3. A atmosfera de romantismo que Luciano propôs se completava com uma trilha sonora que começava com “*What a wonderful world*” e acabava na viagem de “*Lucy in the sky with diamonds*”.
4. João Gilberto em São Paulo Quando: 14 e 15 de agosto (quinta e sexta), às 21h Onde: Auditório Ibirapuera, Av. Pedro Alvares Cabral, portão 03 Quanto: Platéia – *R$ 360,00 (inteira)* – R$ 180,00 (meia)/Setor Superior fileiras de M a P – *R$ 30,00 (inteira)* – R$ 15,00 (meia)
5. Gestor ambiental ‘*administra*’ o meio ambiente

Figure 2.6: Several examples in which quotation marks are not used to delimit quotations

In Figure 2.5, we present examples of quotations, in *italic*, in a given text. We see by those examples quotations usually follow a specific formation pattern, i.e., they are between quotation marks **’ ” -**. By those examples, we could think this task is easily solved by a rule based system.

However, there are many situations a quotation mark is not used to delimit quotations as we present in Figure 2.6. The mispredicted quotations are in *italic*. In the first example, the slash is not used to start a quotation, but to link the verb *aproveitar*[1] to the reflexive pronoun *se*[2]. A soccer stadium

[1] take advantage of

[2] itself

nickname is between quotes in the second example. In the third example, the two expressions between quotes are song names. Slashes are used as commas in the fourth example. In the fifth example, the word between quotes is a pun about the act of an environmental manager[3] manage[4] the environment[5]. From those examples, we see that the quotation identification subtask is not trivial.

1. ”*Ou a gente vai para perto ou então chama os amigos para casa mesmo*”$_1$, diz **Anna Sofia**$_1$.
2. **Mantega**$_1$ se diz ‘*satisfeito*’$_1$ com comportamento da inflação
3. –*Vou entrar com muita determinação. Estou há alguns jogos fora, e essa é a hora de mostrar meu valor e permanecer na equipe*$_1$ – diz **ele**$_1$, que terá a chance por causa da suspensão de **Triguinho**$_2$.
4. **GloboEsporte.com**$_1$: *O Garrincha foi reprovado no exame psicológico, mas acabou sendo um dos destaques da Copa de 1958. O senhor pensou em vetá-lo da competição?*$_1$ **João Havelange**$_2$: *Ele era um homem especial, que veio da roça. (...) Era um fenômeno. (...) Toda regra tem sua exceção.*$_2$

Figure 2.7: Several examples of association between quotation and author

In Figure 2.7, we present examples of association between quotation and author. By those examples, we could think the best solution for a rule-based system would be to link the quotations to the nearest coreferences.

1. **O jogador**$_1$, que será titular contra o **Atlético-PR**$_2$, neste domingo, fora de casa, pelo Campeonato Brasileiro, prefere ver o fato pelo lado otimista, tendo o grupo como pensamento principal.–*O nosso treinador sempre frisa que o Botafogo não é 11 jogadores. O time que deseja ser campeão precisa ter 25 atletas de nível, e acho que o nosso grupo tem essa característica*$_1$ – observa.
2. Sempre que termina **eu**$_1$ brinco com a **Gabi**$_2$: ‘*Estou ralado... mas eles estão muito mais*’$_1$.
3. Segundo o Evangelho de **Mateus**$_1$, **Pedro**$_2$ teria dado mostras impressionantes da fé em seu mestre ao declarar a **ele**$_3$: ‘*Tu és o Cristo, o Filho do Deus vivo*’$_2$.

Figure 2.8: Several examples of association between quotation and author in which the author is not the nearest coreference

[3] gestor ambiental
[4] administrar
[5] meio ambiente

However, in many cases, the quotation author is not the nearest coreference as we present in Figure 2.8.

By those examples, we see the difficulties of the Quotation Extraction task. A rule based system would need many rules to treat non trivial cases of quotation identification and association between quotation and author. Furthermore, it is very time-consuming to create rules which deal with all different cases.

Table 2.1: Examples of the several quotation types found in news

Quotation Type	*Source*
Direct	GLOBOESPORTE.COM: Qual a importância da conquista da Copa do Mundo de 1958?
Direct	‘As colunas da entrada também foram atingidas’, disse ela, em entrevista ao G1, por telefone.
Mixed	O Copom disse ainda acreditar que a atual postura de ‘política monetária’ [subida dos juros], a ser mantida ‘enquanto for necessário’, irá assegurar a convergência da inflação para a trajetória das metas.
Indirect	Lula ressaltou que o Brasil se consolidou como o principal parceiro da Áustria na América do Sul.

Quotations are divided into types (34) as we present them in Table 2.1. Our proposal intends to identify all types of direct and mixed quotations.

2.2 Related Work

Quotation Extraction has been previously approached using different techniques and for several languages. The *NewsExplorer*[6] system extracts quotations from multilingual news (30). It uses lists of verbs of speech, e.g. *said*, *commented*, quotation marks, e.g. ‘ “, general modifiers, e.g. *yesterday*, determiners, e.g. *the*, and people’s names. Then, it uses regular expressions to identify quotations and to associate them to their authors. NewsExplorer does not detect quotations associated to anaphoras.

The *Sapiens* system extracts quotations from news wires in French (5). It identifies quotation candidates selecting all parts of text between quotation marks. Then, it processes the text using a deep syntatic parser and later on looks for specific constructions used in reported speech. When it finds a match, it verifies if the verb head of the main clause belongs to a list of 114 verbs of speech. If it is true, the quotation is identified and the author is associated

[6] http://press.jrc.it/NewsExplorer

to it. If the author is an anaphora, it is solved using a module of anaphora resolution.

The VERBATIM[7] system extracts quotations for Portuguese (34). It uses 19 regular expressions and a list of 35 verbs of speech to identify quotations and their authors. VERBATIM does not detect quotations associated to anaphoras.

The EVRI[8] portal offers a Quotation Extraction API for English news feeds (21). It uses automatic annotation for part-of-speech, phrase type, grammatical role, named entity and coreference. It determines quotation candidates based on a list of verbs of speech and quotation marks. Then, it uses a rule-based approach to identify the best quotation candidates. Finally, it associates the subject of the verb of speech to each identified quotation.

Our proposal differs from previous work since we use ML to automatically build specialized rules instead of human-derived rules. While ML algorithms build models with strong generalization power, human-derived models generally present a lack of generalization. Thus, even small changes in the writing style may need big modifications in the human-derived rule set. In addition, we are able to easily adapt our model to other languages, needing only a list of verbs of speech for a given language. The previously proposed systems would probably need a rule set adaptation to correctly classify the quotations, which would be time consuming.

[7] http://irlab.fe.up.pt/p/verbatim

[8] http://www.evri.com