Source Texts Annotation for Rewriting References to People in Automatic Multi-Document Extracts

Luana Fonseca Cristini[1,2] and Ariani Di Felippo[1,3]

[1] Interinstitutional Center for Computational Linguistics (NILC), São Carlos/SP, Brazil
[2] College of Letters and Sciences (FCL), São Paulo State University (UNESP)
Rodovia Araraquara-Jaú Km 1, Araraquara, 14800-901, Brazil
[3] Language and Literature Department (DL), Federal University of São Carlos (UFSCar)
Rodovia Washington Luís, km 235 - SP 310, São Carlos, 13565-905, Brazil
{luanafcristini;arianidf}@gmail.com

Abstract. Referring expressions in automatic multi-document extracts can be problematic since the sentences extracted from different source texts might contain too little, too much, or repeated information about the referent. In order to improve the cohesion of such extracts, Di Felippo (2016) described the co-reference chains in the human multi-document summaries of CSTNews (i.e., a reference corpus for Automatic Summarization in Portuguese), and proposed a set of rewrite rules for references to people. To evaluate the impact of the rules on automatic extracts, we carried out the annotation of the co-reference chains with mentions to people in the source texts of CSTNews, since they are necessary to apply the rules. In this paper, we emphasize the guidelines, the graphical tool and the statistical results of the annotation.

Keywords: Multi-document Summary, Corpus Annotation, Co-reference Chain.

1 Introduction

Multi-Document Summarization (MDS) is a computational application responsible for generating a single summary from a collection of documents dealing with the same topic [1]. MDS has been extensively investigated for allowing fast and efficient access to information of interest and improving other Natural Language Processing (NLP) applications [2].

Despite the notable advance of MDS over the last decade, automatic extracts still present several problems of cohesion and coherence that affect its informativeness and linguistic quality. Some of the problems are related to references to named entities (e.g., persons, organizations, and locations), such as the occurrence of “first mention (to an entity) without explanation”, “subsequent mention with explanation”, “acronym without explanation” and others [3][4][5][6].

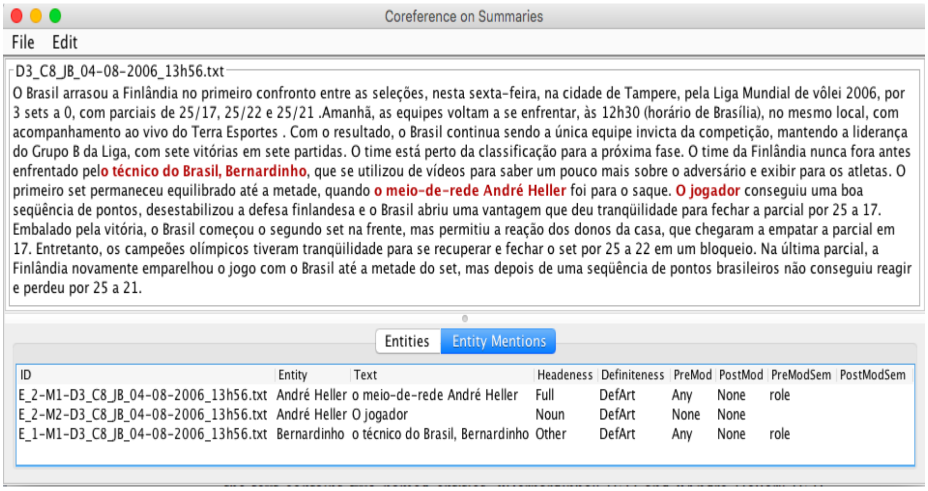
To address these problems, some authors have showed that reference rewriting (as a post-editing MDS step) helps improve automatic extracts [7][8]. Since such rewrites are based on language-dependent rules, [9] first described the co-reference chains to people and organization in news multi-document abstracts writing by humans. The 100-words abstracts were manually built for each collection of the CSTNews corpus [10]. The author used the characterization to develop rules for rewriting first and subsequent references to these types of named entities. The rewrite rules for references to people, for example, ensure that the first mention is descriptive, containing the full name and a pre-modification, whereas subsequent mentions are as brief as possible, using only the *first name* or a common noun that usually indicates the role of the person . Moreover, the rules stablish that the adequate references, which satisfy the rule's conditions, should be selected from the input texts (or collection) of the automatic extracts. Since out next step is to evaluate the impact of the rules on the quality of news extracts, we annotated the co-reference chains with mentions to people in the source texts of CSTNews. In this paper, we describe such task.

In Section 2, we briefly describe the corpus and focus on the annotation procedures. In Section 3, we present the statistics of the annotation. Finally, in Section 4, our future steps and some final remarks will be given.

2 The Annotation of References to People in CSTNews

We used the CSTNews [10], a reference corpus for Automatic Summarization in Brazilian Portuguese. The corpus comprises 50 clusters of news texts from a range of categories (i.e., sports, world, economy, domestic politics, science, and daily news), containing a total of 140 texts. Each cluster contains 2 or 3 news texts on the same topic from different journalist sources (*Folha de São Paulo, Estadão, O Globo, Jornal do Brasil,* and *Gazeta do Povo*), with 42 sentences on average.

In order to annotate the source texts of CSTNews, we have used almost the same graphical tool and guidelines as [9]. Thus, we carried out a semi-automatic corpus annotation using MCATool (*Multi-document Coreference Annotation Tool*), an easy-to-use tool adapted for this mono-document annotation task. Its main window is broken into two parts, the "text pane" on top, and the "editor pane" on bottom (Figure 1).



D3_C8_JB_04-08-2006_13h56.txt

O Brasil arrasou a Finlândia no primeiro confronto entre as seleções, nesta sexta-feira, na cidade de Tampere, pela Liga Mundial de vôlei 2006, por 3 sets a 0, com parciais de 25/17, 25/22 e 25/21 .Amanhã, as equipes voltam a se enfrentar, às 12h30 (horário de Brasília), no mesmo local, com acompanhamento ao vivo do Terra Esportes . Com o resultado, o Brasil continua sendo a única equipe invicta da competição, mantendo a liderança do Grupo B da Liga, com sete vitórias em sete partidas. O time está perto da classificação para a próxima fase. O time da Finlândia nunca fora antes enfrentado pelo técnico do Brasil, Bernardinho, que se utilizou de vídeos para saber um pouco mais sobre o adversário e exibir para os atletas. O primeiro set permaneceu equilibrado até a metade, quando o meio-de-rede André Heller foi para o saque. O jogador conseguiu uma boa seqüência de pontos, desestabilizou a defesa finlandesa e o Brasil abriu uma vantagem que deu tranqüilidade para fechar a parcial por 25 a 17. Embalado pela vitória, o Brasil começou o segundo set na frente, mas permitiu a reação dos donos da casa, que chegaram a empatar a parcial em 17. Entretanto, os campeões olímpicos tiveram tranqüilidade para se recuperar e fechar o set por 25 a 22 em um bloqueio. Na última parcial, a Finlândia novamente emparelhou o jogo com o Brasil até a metade do set, mas depois de uma seqüência de pontos brasileiros não conseguiu reagir e perdeu por 25 a 21.

ID	Entity	Text	Headeness	Definiteness	PreMod	PostMod	PreModSem	PostModSem
E_2-M1-D3_C8_JB_04-08-2006_13h56.txt	André Heller	o meio-de-rede André Heller	Full	DefArt	Any	None	role	
E_2-M2-D3_C8_JB_04-08-2006_13h56.txt	André Heller	O jogador	Noun	DefArt	None	None		
E_1-M1-D3_C8_JB_04-08-2006_13h56.txt	Bernardinho	o técnico do Brasil, Bernardinho	Other	DefArt	Any	None	role	

Figure 1. MCATool main interface and the annotation of a CSTNews source text.

The editor pane is generally customized for the annotation task at hand, while the main text pane is expected to remain constant. The texts in the main text pane cannot be edited, only annotated. At the editor pane, there are two distinct sheets that can be displayed in the same space by selecting the appropriate tab indicating their contents: Entities and Entity Mentions. In Figure 1, the text pane exhibits the text and the editor pane shows the sheet content of the tab Entity Mentions. MCATool usage is simple – it allows for three kinds of operations: adding new mention for new entity, adding new mention, and describing the mention attributes. The list of attributes should be defined in an annotation configuration file, which we load before annotation.

Based on [9], we limited person entities only to humans and single individuals, and the extent of a mention to the entire nominal phrase (NP), including determiners and modifiers. However, only NPs with mentions to people as head were tagged, since non-head mentions have less variation in terms of modification and name realization [11]. Thus, mentions in a modification function or possessive construction (in bold) (i.e., [the **Bush** administration] and [**Hitler**'s yacht]) were not annotated. Moreover, we focused on the annotation of chains with entities that were mentioned *by name* at least one time in a text (in bold). This is the case of the entity André Heller in Figure 1, whose first mention conveys the person´s name (i.e., *o meio-de-rede **André Heller*** ("the middle blocker **André Heller**"). Therefore, there is no chain in our annotated data composed by common noun phrases only (e.g., [*o meio-de-rede* > *o jogador*]).

We described each mention according to 3 features: *headedness*, *modification*, and *definiteness* [9]. The attribute headedness could be annotated with one of 6 values: *full name*, *last name*, *first name*, *common noun*, *pronoun* and *other* (nickname). Modification encompasses the occurrence of pre- and post-modification. The pre-modifiers are adjectives or nouns that precede a NP head, indicating: (i) affiliation; (ii) title, (iii) role, or (iv) temporal role (adjectives such as "former" or "interim"). We tagged mentions with one of these pre-modifier as containing *any* pre-modification. The post-modification was characterized by the occurrence of an *appositional phrase* (AppP), *prepositional phrase* (PrepP), *adjectival phrase* (AdjP), *relative clause* (RC) or *other* (parenthetical observations) after the NP head. If a mention did not have modification, the pre- and post-modification attributes were given the value *none*. For definiteness, we used the following values: *definite article*, *indefinite article*, *possessive*, and *none*.

In Figure 1, we see the annotation of the co-reference chains with mentions to people in Document 3 (D3) of *cluster* C8, which was compiled from *Jornal do Brasil* (JB) (04 Ago 2006). Specifically, the text contains two named people, *Bernardinho* (E1) and *André Heller* (E2). The unique mention to E1, *o técnico do Brasil, Bernardinho* ("the Brazilian coach, Bernardinho") (E1-M1), was described with the following attributes and values: headedness=other (*Bernardinho*), definiteness=artdef (*o*) ("the"), pre-modification=any (role) (*técnico do Brasil*), and post-modification=none. The E2 occurs 2 times in the text, thus the correspondent chain has 2 mentions. The first mention to E2 (E2-M1), *o meio-de-rede André Heller*, has almost the same attribute and values as E1-M1. The exception is headedness, which value is full name. The second mention to E2 (E2-M2), *O jagodor*, was tagged with the following attribute-value pairs: headedness=noun (*jogador*), definiteness=defart (*O*), pre-modification=none, and post-modification-none.

Specifically, each news text of CSTNews was thus semi-automatically annotated by 1 linguist (i.e., one of the paper´s authors) with the support of MACTool. The human annotator went through a training phase first, in which a general study of guidelines and editors for co-reference annotation was carried out during approximately 2 weeks. After training, the annotation started. The annotation itself has taken about 1 month in a daily one-hour section basis.

3 Results

From the 140 news in CSTNews, we computed 119 texts with named references to people. Out of the 50 collections on the corpus, 3 have no references to people: C11, C22, and C30. The no-occurrence of references to named people in such clusters seems to be associated with their topics, since C11 and C22 are about daily news ("a number of criminal attacks" and "bad weather closes an airport", respectively) and C30 is about economy ("annual profits of a bank"). Specifically, we annotated 506 co-reference chains with 1073 mentions. This means that, on average, each text in CSTNews contains 4.2 chains (506/119), and such chains have an average of 2.1 mentions (1073/506). Thus, one may see that the texts in our corpus have few references to people and the chains are not too long. This might be due to the relatively short length of the news (42 sentences on average). We also identified 323 distinct people mentioned by name in the 119 texts, which means that, on average, there are 2.7 distinct entities per text (323/119). In addition, we highlight that the xml files generated by MCATool, containing the source text upon which the annotation was performed as well as its annotation, will be available soon at the Sustento[1] webpage.

4 Future works and Final Remarks

In the following steps, we will first use the GistSumm [12] and RSumm [13][14] summarizers to generate automatic multi-document extracts (with 110 words each) for the CSTNews clusters, and then we will manually rewrite problematic references in such extracts based on the rules of [9]. GistSumm follows a very naive approach, by simply juxtaposing all the texts of the input and selecting the sentences according to the frequency of their words. RSumm, which is based on the relationship maps proposed by [15], it is a state-of-art multi-document summarizer for MDS of texts written in Portuguese. Thus, we intend to verify the impact of the rules on the overall extracts generated by these two summarizers that have different performances. Our evaluation will be manual by comparing an automatic extract and its rewritten variant according to the guidelines proposed by [16]. We expect that the evaluation of the rules [9] – which will be just possible due to the co-reference annotation described here – reveals that the revision of references in extracts is a potential strategy to shift away from the purely extractive MDS in Portuguese to a partially abstractive one.

[1] Sustunto is a research project which aims at generating linguistic knowledge for MDS in Brazilian Portuguese (http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23).

Acknowledgments. The authors are grateful to FAPESP for supporting this work.

References

1. Mani, I. Automatic summarization. Amsterdam: John Benjamins Publishing Co. (2001).
2. Nenkova A.; Mckeown. K. Automatic summarization. Foundations and Trends in Information Retrieval, 5(2-3), 103–233 (2011).
3. Otterbacher, J. C.; Radev, D. R.; Luo, A. Revisions that improve cohesion in multi-document Summaries: A Preliminary Study. In: Proceedings of the Workshop on Automatic Summarization (ACL-02), Philadelphia/USA, pp. 27–36 (2002).
4. Kaspersson, T.; Smith, C.; Danielsson, H.; Jönsson, A. This also affects the context - errors in extraction based summaries. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul/Turkey, pp. 173–178 (2012).
5. Friedrich, A., Valeeva, M., Palmer, A. LQVSumm: a corpus of linguistic quality violations in multi-document summarization. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik/Iceland, pp. 1591–1599 (2014).
6. Dang, H.T. Overview of DUC 2005. In: Proceedings of the Document Understanding Conference (HLT/EMNLP Workshop on Text Summarization) (2005).
7. Nenkova, A.; Siddharthan, A.; McKeown, K. Automatically learning cognitive status for multi-document summarization of newswire. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver/Canada, 241–248 (2005).
8. Siddharthan, A., Nenkova, A., McKeown, K. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. Computational Linguistics 37(4), 811–842 (2011).
9. Di-Felippo, A. Revisão de sumários baseada em conhecimento: transformando extratos multidocumento em *abstracts*. Relatório de Bolsa de Pesquisa no Exterior (BPE) (FAPESP #2015/01450-5). Disponível em http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23.
10. Cardoso, P. C. F.; Maziero, E. G.; Jorge, M. L. C.; Seno, E. M. R.; Di-Felippo, A.; Rino, L. H. M.; Nunes, M. G. V.; Pardo, T. A. S. CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In: Proceedings of the 3rd RST Brazilian Meeting, Cuiabá/MT/Brazil (2011)
11. Nenkova, A.; McKeown, K. Improving the Coherence of Multi-document Summaries: a Corpus Study for Modeling the Syntactic Realization of Entities, Columbia University, CS Department Technical Report, CUCS-001-03, 2003.
12. Pardo, T. A. S. GistSumm - GIST SUMMarizer: Extensões e novas funcionalidades, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (2005)
13. Ribaldo, R.; Akabane, A. T.; Rino, L. H. M.; Pardo, T. A. S. Graph-based Methods for Multidocument Summarization: Exploring Relationship Maps. Complex Networks and Discourse Information. In: Proceedings of the 10th International Conference on Computational Processing of Portuguese (LNAI 7243), Coimbra/Portugal, pp. 260–271 (2012)
14. Ribaldo, R. Investigação de Mapas de Relacionamento para Sumarização Multidocumento. Monografia de Conclusão de Curso. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 61p. (2013)
15. Salton, G.; Singhal, A.; Mitra, M.; Buckley, C. Automatic Text Structuring and Summarization. Information Processing & Management 33(2), 193–207 (1997)
16. Dang, H. Overview of DUC 2005. In Proceedings of the Document Understanding Conference, Vancouver, B.C., Canada (2005).