Descrição formal na identificação de Redundância na Sumarização Automática Multidocumento

Jackson Souza
UFSCar

O acesso e a disponibilização da informação digital têm aumentado esponencialmente nos últimos anos. De acordo com a Cisco-Visual-Networking-Index (2017), em 2016 foram produzidos 1 Zettabytes de informação, fato que não aconteceu desde a criação da Internet. Diante desse cenário, subáreas do Processamento Automático de Línguas Naturais (PLN), em especial, a Sumarização Automática Multidocumento (SAM), visam produzir soluções computacionais que consigam lidar com essa enorme quantidade de dados.

A SAM tem ganhado visibilidade nesse contexto porque objetiva automatizar a produção de sumários (ou resumos) baseados em coleções de textos advindos de fontes distintas que abordem um mesmo assunto (MANI, 2001). Por essa tarefa ser de tal natureza, é necessário que sejam elencadas quais sentenças de cada um dos textos-fonte evidenciam o tópico principal do evento veiculado. Nesse processo, evita-se que haja redundância e contradição entre o conteúdo selecionado, além de detectar se há informações complementares entre as sentenças para compor o sumário. Configuram-se, dessa forma, os três principais fenômenos multidocumento na área de SAM: Redundância, Contradição e Complementaridade, respectivamente (MAZIERO, 2010).

A identificação desses fenômenos têm sido majoritariamente realizadas de maneira automática com base em características com pouco conhecimento linguístico, como quantidade de palavras em comum entre pares de sentenças, (MAZIERO, 2010). Para o Português do Brasil (PB), há esforços de caracterizar tais fenômenos considerando características linguísticas específicas (p.ex. SOUZA *et. al.*, 2012).

Em especial, a Redundância tem sido caracterizada linguisticamente com base em traços linguísticos na superfície do texto, evidenciando a presença ou a ausência de informações em comum entre pares de sentenças, como exemplificado em (1).

(1)
S1: A margem de erro é de 2 pontos porcentuais.
S2: A margem de erro é de dois pontos percentuais, para mais ou para menos.

Em (1), a informação veiculada na primeira sentença é a mesma da segunda, tendo somente uma variação lexical (numeral por extenso e algarismo, e "*porcertual*" por "*percentual*"). Além disso, S2 apresenta informação extra com relação a S1 (*"para mais ou para menos"*). De acordo com os métodos vigentes, seria facilmente detectávelde forma automática uma alta redundância entre os pares de sentença, já que a variação da forma (palavras) não prejudica a verificação do conteúdo (informação).

Entretanto, há casos em que há mudanças consideráveis quanto à forma que, de

maneira automática, não é capaz de apontar com maior precisão a Redundância presente em um par de sentenças, como exemplificado em (2).

(2)
S1: A TAM anunciou o cancelamento de 68 vôos nesta terça (24) e o remanejamento de outros 22.
S2: A TAM cancelou nesta terça-feira 68 vôos e desviou 22 para Guarulhos.

Em (2), o par de sentenças veicula a informação sobre o cancelamento de vôos da empresa TAM em aeroportos de São Paulo. A mudança de forma entre os pares de sentenças (p.ex.: "*A TAM anunciou o cancelamento"*, em S1, e *"A TAM cancelou"*, em S2), faz com que os métodos automáticos de identificação apontem que há pouco conhecimento compartilhado entre o par. No entanto, sabe-se que ambas as sentenças veiculam o mesmo conteúdo, em detrimento da variação da forma.

Objetiva-se, então, neste trabalho, utilizar os métodos de Cálculo Lambda, os quais permitem representar formalmente o sentido (neste caso, a informação) compartilhada entre as sentenças de um par. O par de sentenças em (1), por exemplo, pode ser representado, respectivamente, por **S1: λm[ser(p)]1** e **S2: λm[ser(p)/λp(mm)]**, sendo as variáveis "*margem de erro*" equivalente a "m", "*de 2 pontos percentuais*", igual a "p" e "*para mais ou para menos*" representado por "mm".

Para o estudo formal do comportamento linguístico da Redundância, foi utilizado o *corpus* multidocumento CSTNews (CARDOSO *et al.*, 2011), cuja característica principal é ter seus textos jornalísticos anotados com os pressupostos da *Cross.document Structure Theory* (CST) (RADEV, 2000) e separados em pares de sentenças que evidenciam as relações da teoria. Fez-se, então, um recorte para selecionar somente os pares de sentença anotados com as relações *Identity, Elaboration, Equivalence, Contradiction, Summary, Subsumption* e *Overlap*, as quais traduzem a Redundância.

Espera-se com esse estudo motivar a descrição formal dos fenômenos multidocumentos da SAM, já que o cenário atual mostra que a caracterização somente pela forma limita a definição do fenômeno, além de haver casos em que não é possível representar o conteúdo veiculado.

REFERÊNCIAS

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, pp. 88-105. Cuiabá/MT, Brasil. 2011.

CISCO-VISUAL-NETWORKING-INDEX. *Forecast and Methodology*. 6 de Junho de 2017. Disponível em: www.cisco.com/c/en/us/solutions/collateral/service-

provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf. Acesso em: 04 de Agosto de 2017.

MANI, I. Automatic Summarization. John Benjamins Publishing Co., Amsterdam. 2001.

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying multi-document relations. In: International Workshop on Natural Language Processing and Cognitive Science. Funchal, Madeira. p. 60-9. 2010.

RADEV, D. R. A common theory of information fusion from multiple text sources, step one: cross-document structure. In: ACL Sigdial Workshop on Discourse and Dialogue, 1, 2000, Hong Kong. Proceedings of ACL Sigdial Workshop on Discourse and Dialogue. Hong Kong, 2000, p. 74-86.

SOUZA, J. W. C.; DI-FELIPPO, A.; PARDO, T. A. S. Investigação de métodos de identificação de redundância para Sumarização Automática Multidocumento. Série de Relatórios do NILC. NILC-TR-12. São Carlos-SP. 2012.

SOUZA, J. W. C. Dissertação de Mestrado. Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos. São Carlos-SP. 2013.