

O Desenvolvimento de um Sistema Computacional de Sumarização Multidocumento com Base em um Método Linguisticamente Motivado

Guilherme Gonçalves, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos/SP, Brasil

guilherme3.goncalves@usp.br, taspardo@icmc.usp.br

Abstract. *This paper presents the studies conducted in the area of Natural Language Processing, more specifically, in Automatic Multi-document Summarization. We describe the steps for the production of a computational prototype, based on a linguistically motivated method, for summarizing news texts in Portuguese.*

Resumo. *Este trabalho apresenta os estudos realizados na área de Processamento de Linguagem Natural, mais especificamente, em Sumarização Automática Multidocumento. São descritos os passos para a produção de um protótipo computacional, baseado em um método linguisticamente motivado, para a produção de sumários de notícias jornalísticas escritas em português.*

1. Introdução

Atualmente, temos acesso a um volume de dados tão grande que, muitas vezes, não conseguimos processar e apreender todas essas informações. Além disso, cada vez menos dispomos de tempo para atividades como essas. Assim, é necessária uma forma eficiente e rápida de assimilar informações. Diante dessa dificuldade, uma ferramenta de sumarização multidocumento se mostra interessante e vantajosa.

A Sumarização Automática Multidocumento (SAM) consiste na produção automática de um único sumário (também chamado resumo) a partir de um grupo de textos sobre um mesmo tópico ou sobre tópicos relacionados (Mani, 2001). A SAM seleciona as informações potencialmente mais relevantes, evitando que se perca tempo, identificando e excluindo informações desnecessárias e menos importantes. Dessa maneira, ela norteia o leitor a saber se o assunto tratado em um texto ou em uma coleção de textos é realmente de seu interesse. Como ilustração, a Figura 1 mostra um exemplo de sumário multidocumento produzido manualmente a partir de dois textos jornalísticos que reportavam problemas de saúde de um ex-jogador de futebol.

Maradona voltou a ter problemas de saúde no fim de semana e foi internado novamente em um hospital em Buenos Aires. Ele teve uma recaída da hepatite aguda. Ele melhorou e está estável, mas continuará internado. Maradona desenvolveu hepatite por excesso de álcool, mas, nesta recaída, ele não estava bebendo e a causa ainda é indeterminada, segundo seu médico.

Figura 1. Exemplo de sumário multidocumento

Os trabalhos na área de Processamento de Linguagem Natural (PLN), área à qual a SAM pertence, geralmente seguem as seguintes etapas: de estudo linguístico, representacional e de implementação. Na etapa de estudos linguísticos, são realizadas as investigações dos métodos e atributos utilizados pelos humanos na tarefa. Já na representacional, ocorre a formalização das estratégias identificadas. Finalmente, na etapa de implementação, ocorre a produção de sistemas computacionais que empregam o que foi definido nas etapas anteriores.

Na SAM, os trabalhos realizados normalmente apresentam os seguintes processos: análise, transformação e síntese. No processo inicial de análise, um ou mais textos-fonte são processados e uma representação interna com todo o conteúdo dos textos é produzida. Essa representação deve ser formal o suficiente para ser processada automaticamente. O processo de transformação realiza a seleção de conteúdo sobre a representação interna dos textos-fonte, produzindo a representação interna do sumário, a qual contém o conteúdo mais importante a ser veiculado textualmente. Por fim, o processo de síntese expressa em língua natural a representação interna do sumário.

Em uma abordagem pioneira, Camargo (2013) estudou os fundamentos linguísticos da SAM para formalizar estratégias de seleção de conteúdo para compor o sumário (no âmbito da etapa de transformação, portanto). Este trabalho utiliza os estudos conduzidos por Camargo, sendo um dos primeiros a utilizar um método linguisticamente motivado para o português brasileiro. A proposta é produzir um sistema computacional que seja capaz de produzir sumários multidocumento informativos e úteis aos humanos.

A seguir, apresentamos os principais trabalhos relacionados. Na Seção 3, descrevemos e apresentamos o estado atual do protótipo implementado. A Seção 4 apresenta algumas considerações finais.

2. Trabalhos Relacionados

Segundo Mani (2001), na SAM, podemos encontrar dois tipos de abordagens: a superficial e a profunda. Na superficial, utilizam-se métodos estatísticos e empíricos, que podem ser modelados por atributos, que ditam quais sentenças devem ir para o sumário, podendo ser, por exemplo, a primeira sentença de um texto-fonte, aquela que contém as palavras mais frequentes ou ainda sentenças que tenham seu tamanho dentro de uma certa faixa pré-determinada. Na abordagem profunda, faz-se uso de mais conhecimento linguístico para se interpretar o conteúdo textual e selecionar as informações que compõem o sumário de forma mais consistente. Por exemplo, pode-se fazer uso de recursos e modelos semânticos e discursivos para relacionar as partes dos vários textos e detectar e tratar apropriadamente os vários fenômenos multidocumento, que incluem as similaridades e diferenças que existem entre os textos a serem sumarizados. Na SAM, um dos modelos discursivos mais utilizados é a CST (*Cross-document Structure Theory*) (Radev, 2000), que prevê que os segmentos textuais de vários textos sobre um mesmo tópico podem apresentar relações como equivalência, contradição, elaboração, subsunção e atribuição, dentre várias outras.

O trabalho de Camargo (2013) considera o fato de que sumários são compostos por informações que apresentam características recorrentes, a ponto de revelar estratégias de sumarização. Em seu trabalho, Camargo utilizou os textos-fonte das 50 coleções do corpus multidocumento para o português brasileiro CSTNews (Cardoso et al., 2011) para realizar sua investigação. Camargo identificou que as sentenças selecionadas para compor

os sumários multidocumento comumente apresentam certos atributos com valores padronizados, como localização das sentenças no texto e redundância. Essa constatação foi confirmada pelo conjunto de regras formais aprendidas por um algoritmo de Aprendizado de Máquina (AM) a partir das caracterizações encontradas em sua investigação. A Figura 2 apresenta algumas das regras geradas pelo algoritmo de AM. Essas regras são aplicadas a cada sentença dos textos-fonte e indicam que sentenças devem ser selecionadas para compor o sumário. Nesse caso, o sumário – chamado extrato – é formado pela simples justaposição de sentenças selecionadas.

<p>Se uma sentença estiver no início de algum texto, ela deve ser selecionada para o sumário.</p> <p>Senão, se ela tiver grande redundância com as demais do texto, ela deve ser selecionada para o sumário.</p> <p>Senão, se a redundância for baixa, mas suas palavras forem frequentes, ela deve ser selecionada para o sumário.</p> <p>Senão, ela não deve ser selecionada para o sumário.</p>
--

Figura 2. Regras geradas pelo algoritmo de AM

A partir da quantidade de acertos de cada regra gerada pelo algoritmo de AM, é possível calcular a precisão que dada regra possui em selecionar uma sentença para o sumário. Essa precisão é calculada por meio da razão do número de acertos pelo total de sentenças analisadas pela regra.

Na SAM, a redundância é um atributo que influencia fortemente na escolha de conteúdo, uma vez que a redundância indica as informações mais importantes presentes nos textos-fonte. Trabalhos anteriormente realizados na área constataram que informações que aparecem muitas vezes são consideradas importantes nos textos. Pelo fato dos autores ficarem retomando as mesmas informações, isso indica que elas devem ser salientadas no sumário correspondente. Dessa maneira, considera-se que informações redundantes expressam o assunto principal dos textos na maior parte das vezes. No entanto, para se compor o sumário, não é interessante que as sentenças selecionadas apresentem redundância. Nesse contexto, trabalhos como o de Souza (2012) investigam métodos de identificação da redundância entre sentenças a serem inseridas em um sumário, procurando formas de evitar a redundância no sumário.

No protótipo desenvolvido neste trabalho, implementam-se as regras de seleção de conteúdo obtidas por Camargo. Em termos de remoção de redundância no sumário, pretende-se, futuramente, utilizar o método desenvolvido por Souza.

3. O Protótipo

Nessa seção, será descrito o trabalho realizado até o presente momento, ou seja, o estado da ferramenta de sumarização que o projeto tem como finalidade.

Inicialmente, o protótipo computacional realiza o pré-processamento dos textos-fonte. Para o pré-processamento, é realizada a leitura dos textos-fonte, que precisam estar no mesmo diretório onde se encontra o arquivo executável do protótipo. Além do executável, é necessário que esteja no mesmo diretório o arquivo que contém a *stoplist*, ou seja, uma lista formada por *stopwords*, palavras que não expressam conteúdo relevante dentro da sentença. Em geral, essas palavras são descartadas durante o processo de

sumarização, pois podem interferir nos resultados.

Para poder realizar tanto a análise morfológica como sintática das sentenças é utilizado o parser PALAVRAS (Bick, 2000). O PALAVRAS apoia-se num léxico de 50.000 lemas e em milhares de regras gramaticais para fornecer uma análise completa. Esse software baseia-se na *Constraint Grammar* (Karlsson et al., 1995). Nesse paradigma, as regras dependentes de contexto são compiladas em uma gramática que atribui etiquetas gramaticais (*grammatical tags*) a palavras ou outros símbolos (*tokens*) em texto.

O PALAVRAS gera para cada texto um arquivo de saída na forma ilustrada pela Figura 3 a seguir. Como se vê na figura, a análise foi feita para a frase “*Mesmo após lei, prazo para tratar câncer ainda é descumprido no país*”. Cada palavra é mostrada em uma linha, acompanhada de suas informações linguísticas. Desse processamento, são extraídas importantes definições sobre as palavras dentro das sentenças. Para a linha: “é [ser] <fmc> <aux> V PR 3S IND VFIN @FS-STA #10->0”, o lema encontra-se entre colchetes, e o V indica que a palavra é um verbo. As definições extraídas são utilizadas para o cálculo de atributos, como tamanho da sentença e frequência das palavras lematizadas, que posteriormente são utilizados para a seleção de conteúdo que formará o sumário.

```
Mesmo [mesmo] <*> <quant> ADV @>P #1->2
após [após] PRP @ADVL> #2->0
lei [lei] <conv> <percep-f> N F S @P< #3->2
$, #4->0
prazo [prazo] <clb> <per> <temp> N M S @SUBJ> #5->10
para [para] <np-close> PRP @N< #6->5
tratar [tratar] <mv> V INF 3S @ICL-P< #7->6
câncer [câncer] <sick-c> N M S @<ACC #8->7
ainda [ainda] ADV @ADVL> #9->11
é [ser] <fmc> <aux> V PR 3S IND VFIN @FS-STA #10->0
descumprido [descumprir] <vH> <mv> V PCP M S @ICL-AUX< #11->10
em [em] <sam-> PRP @<ADVL #12->11
o [o] <-sam> <artd> DET M S @>N #13->14
país [país] <Lciv> N M S @P< #14->12
$. #15->0
```

Figura 3. Modelo de saída do PALAVRAS

O processamento dos arquivos de saída do PALAVRAS é realizado representando as sentenças e palavras por meio de listas encadeadas. Listas encadeadas se mostraram úteis pelo fato de não apresentarem limites prévios de elementos, pois o número de sentenças de um texto para o outro varia muito e o mesmo acontece com as palavras de sentença para sentença dentro de um mesmo texto. Assim, a estrutura de dados ficou como uma lista de textos-fonte, onde cada texto-fonte possui uma lista de sentenças e cada sentença possui sua lista de palavras, e cada palavra é acompanhada de suas informações linguísticas.

O próximo passo foi calcular os atributos segundo o trabalho de Camargo (2013). Para cada atributo, há uma sub-rotina que percorre a estrutura de dados, realizando a

contagem para cada sentença e definindo o valor do atributo em questão para a mesma. Ao final, normalizam-se os valores para que discrepâncias nos dados sejam evitadas. Essas discrepâncias advêm das diferenças de tamanho entre sentenças e textos. Sendo assim, a normalização garante um tratamento mais uniforme dos dados.

Para os atributos profundos do método de Camargo, que são fornecidos pela teoria CST (Radev, 2000), foi necessário realizar a busca em documentos que continham a catalogação das relações para os textos. Tal catalogação foi realizada anteriormente por estudiosos da área de linguística computacional. As relações da CST são responsáveis por fornecer, por exemplo, a informação sobre a redundância das sentenças, que, como já foi dito, é um atributo muito importante na SAM. Esses valores também são normalizados.

Em seguida, implementou-se o conjunto de regras que norteiam a seleção de conteúdo. Essas regras são as mesmas definidas por Camargo em sua investigação. Para tanto, foi implementada uma sequência de comandos de decisão do tipo *if-then-else* para poder definir se a sentença seria selecionada para o sumário. A seleção é feita com base aos valores dos atributos calculados e previamente definidos. Ao selecionar a sentença, essa recebe a precisão da regra que a selecionou. As sentenças selecionadas são, ao fim, ranqueadas pela precisão associadas a elas, sendo que as sentenças mais bem ranqueadas têm preferência para compor o sumário, que tem seu tamanho limitado pela taxa de compressão (especificada em uma porcentagem com base no número de palavras dos textos).

Atualmente, já na fase final de desenvolvimento do protótipo, trabalha-se no método de remoção de redundância para finalmente compor o sumário. Pretende-se, então, avaliar o sistema. Essa avaliação será realizada aplicando-se a medida de avaliação ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin e Hovy, 2003), fortemente empregada na área de sumarização. A medida ROUGE realiza a comparação do sumário automático com o sumário humano, por meio da co-ocorrência de n-gramas, que consiste em verificar a média de quantas vezes cada conjunto de n palavras adjacentes dos sumários automáticos se repetem nos sumários humanos. Essa medida indica a informatividade dos sumários automáticos, sendo este critério um dos mais importantes na área. Essa avaliação deve ser realizada com os sumários humanos que acompanham o corpus CSTNews.

4. Considerações Finais

Neste trabalho, está se estudando um método que potencializa a eficiência da seleção de conteúdo para a SAM. Acredita-se que bons resultados podem ser gerados para a área de pesquisa.

Esse trabalho possui a limitação de depender da catalogação prévia das relações CST. No entanto, trabalhos futuros poderão realizar essa tarefa de forma automática. Há sistemas disponíveis para isso, mas sua acurácia ainda é relativamente baixa, dada a subjetividade desse tipo de análise.

Assim que o protótipo de SAM estiver com um funcionamento razoável, ele será disponibilizado para utilização pública.

Agradecimentos

À FAPESP, ao ICMC e à Pró-reitoria de Pesquisa pelo apoio a este trabalho.

Referências

- Bick, E. (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.
- Cardoso, P.C.F.; Castro Jorge, M.L.R.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews – A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the RST Brazilian Meeting*, pp. 88-105.
- Camargo, R.T. (2013). *Investigação de estratégias de sumarização humana multidocumento*. Dissertação de Mestrado. Universidade Federal de São Carlos.
- Karlsson, F.; Voutilainen, A.; Heikkilä, J.; Anttila, A. (1995). *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter.
- Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In the *Proceedings of the Language Technology Conference*, pp. 71-78.
- Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co. Amsterdam.
- Radev, D. R. (2000). A common theory of information fusion from multiple text sources, step one: cross-document structure. In the *Proceedings of the 1st ACL Workshop on Discourse and Dialogue*, pp. 74-83.
- Souza, J.W.C. (2012). *Investigação de métodos de identificação de redundância para Sumarização Multidocumento*. Trabalho de Conclusão de Curso. Universidade Federal de São Carlos.