MODELO PARA RESUMOS ABSTRATOS DA METODOLOGIA DE ARTIGOS CIENTÍFICOS

Weverson da Silva Pereira[1], Paulo Sérgio Silva Rodrigues[2]
[1,2] Ciência da Computação, Centro Universitário FEI
unifwpereira@fei.edu.br psergio@fei.edu.br

Resumo: Atualmente, devido ao enorme volume de informações científicas, o acesso aos dados digitais tornou-se um dos maiores desafios. Um desses desafios é a automatização da construção de resumos de artigos acadêmicos. Este trabalho tem por objetivo a utilização de processamento de linguagem natural para a condensação da metodologia presente em trabalhos científicos publicados e disponibilizados na internet.

1. Introdução

Com o desenvolvimento da internet e o barateamento do hardware, o volume de informações encontrada hoje em dia pelo cidadão comum é extraordinário. Uma das áreas que mais se beneficiou e que também é grande produtora de informação é a área científica, sobretudo a tecnológica. Sites como IEEE Xplore, Science Direct e ArXiv são exemplos bem conhecidos que disponibilizam praticamente a maior parte do conhecimento desenvolvido pela sociedade.

Entretanto, apesar do volume de informações, a sua organização para leitura é um desafio em paralelo, em que o pesquisador interessado ainda deve criar requisitos para ordenar os dados obtidos de modo a iniciar um levantamento de informação preciso e seguro, sem correr risco de perda de informação.

Uma maneira comum entre pesquisadores para categorizar artigos científicos é segmentar cada artigo em elementos como objetivo geral, metodologia para se alcançar aquele objetivo, resultados obtidos e próximos caminhos a serem tomados a partir de sua conclusão. Embora a variedade de escrita de artigos seja enorme, é por norma padrão que esses quatro elementos devem coexistir de forma consensual na maioria dos artigos publicados, de modo a facilitar o acesso, a leitura e o entendimento de cada trabalho. Pode-se argumentar, no entanto, que esses elementos possuem como base tanto o conhecimento sintático quanto semântico do texto, conduzindo assim o problema para a interpretação de linguagem natural.

A área de processamento de linguagem natural apresentou um avanço considerável em décadas recentes, tornando possível o aperfeiçoamento de técnicas como: resumo e classificação automática de texto, assistentes virtuais, tradução de idiomas entre outros.

Um exemplo é o trabalho de [1], que mostra uma abordagem para resumo abstrato de frases. Essa abordagem, chamada de *Attention-Based Summarization* (ABS), combina um modelo de encoder com um algoritmo para geração de textos, o *beam search decoder*. Utilizando o conjunto de dados de larga escala DUC-2004, o ABS superou todos os modelos considerados estado da arte na época.

O modelo proposto por [2] pode ser visto como uma extensão do modelo de [1], citado acima, tentando solucionar o mesmo problema, mas utilizando uma rede neural recorrente condicional. Este condicionamento é fornecido por um *attention-based convolucional encoder*, impactando no *decoder* para que apenas se concentre nas palavras de entrada apropriadas em cada etapa da geração. Os experimentos mostraram que este modelo ultrapassou em pelo menos 0.9 o estado da arte da época (ABS+) na base de dados Gigaword e utilizando a métrica ROUGE.

Por sua vez em [3], resolveu o mesmo problema usando *attentional encoder-decoder RNN*, mas tratando alguns empecilhos do resumo de texto, como o modelamento de palavras-chave, a captura da hierarquia na estrutura de sentença para cada palavra e a emissão palavras raras ou invisíveis no momento do treinamento. O modelo superou todos estados da arte da época em duas bases de dados diferentes (Gigaword e DUC).

Um passo à frente foi dado por [4], que buscava um modelo que tivesse um bom desempenho ao lidar com resumos de longos textos (artigos científicos, por exemplo). Para isso, foi usado um *hierarchical encoder* para capturar a estrutura do discurso do documento e um *discourse-aware decoder* para gerar o resumo. Por não existir uma base de dados que tivesse longos documentos, eles introduziram bases com artigos coletados de grandes repositórios científicos: ArXiv e PubMed. Os resultados empíricos nessas bases mostraram que o modelo proposto supera significativamente os modelos na métrica ROUGE em comparação com trabalhos posteriores.

Sendo assim, o projeto aqui proposto tem por objetivo a utilização de aprendizado de máquina e mineração de dados para o processamento de linguagem natural apresentada na metodologia de artigos acadêmicos-científicos, de modo a sintetizar as ações realizadas, as técnicas utilizadas e a forma de análise do problema adotada pelos autores; contribuindo assim para a organização desse tipo de informação e para facilitar a sua leitura

2. Recurrent Neural Netowrk (RNN)

Assim como trabalhos anteriores usaram, esse projeto utilizará de uma Rede Neural Recursiva para o resumo automático de texto. Tal escolha se deve ao fato dela ter apresentado uma boa performance nas tarefas de processamento de linguagem natural e à sua capacidade de levar em consideração a posição das palavras nas frases e produzir agrupamentos contextuais de palavras.

RNNs são projetadas para interpretar informações temporais ou sequenciais e possuem a capacidade de usar outros pontos de dados em uma sequência para

fazer melhores previsões. Isso é realizado ao receber entradas e reutilizando as ativações de nós anteriores ou posteriores na sequência para influenciar a saída. Essa rede neural contém *loops* que permitem a persistência da informação, como mostra a figura 1.

Contudo, existe um tipo de *RNN* ainda mais específio, a *Long Short Term Memory (LSTM)*. Esse tipo especial de rede neural recorrente torna mais fácil a recordação de informações antigas na memória e resolve uma dificuldade da RNN que é lidar com grandes sequencias de entrada, problema esse chamado de gradiente de fuga, que pode fazer com que se perca informações importantes de um texto, por exemplo. A LSTM tem "portões" que podem aprender qual informação é importante e quais devem ser descartadas, abordagem esta que poderá ser utilizada nesse projeto.

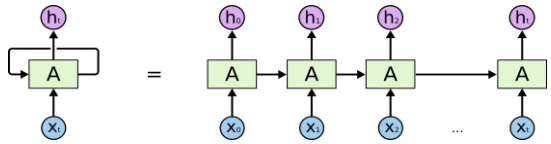

Figura 1 – Exemplo de rede neural em que x(t) representa as entradas e h(t) as saídas. A saída h(1) retornaria uma informação tendo h(0) e x(1) como entradas, por exemplo. Fazendo-a se lembrar do contexto enquanto treina.

3. Base dados utilizadas

Partindo da base de dados fornecida por [4], foi escolhida a base de dados arXiv por ser a menos enviesada quando comparada à Pubmed que tem como objetivo armazenar artigos biomédicos e relacionados à ciências da vida.

Iremos então filtrá-la para que contenha apenas artigos científicos com uma seção exclusiva para a metodologia. A tabela I mostra um comparativo da base extraída com a original.

Após isso, iremos extrair manualmente da seção resumo o segmento que corresponde ao sumário da metodologia. A figura 2 mostra um comparativo da base original e da nova base com foco na metodologia.

```
{                                          {
  'article_id': str,                         'article_id': str,
  'abstract_text': List[str],                'methodology': List[str],
  'article_text': List[str],                 'method_summary': List[str]
  'section_names': List[str],              }
  'sections': List[List[str]]
}
```

Figura 1 – Exemplo de artigo na base de dados utilizada. À esquerda temos um da base de dados original e à direita a base criada.

Tabela I – Estatísticas da base de dados

Base de Dados	Quantidade Docs.	Média de palavras por documento	Média de palavras do resumo
arXiv (original)	215913	4938	220
arXiv (este trabalho)	5719	1335	55

4. Resultados Parciais

Seguindo o método de [4], que propõe o uso de RNNs com *hierarchical encoder* e *attentive discourse-aware decoder* para a sumarização de documentos científicos, obtivemos resumos similar ao de humanos. Nota-se na tabela II que os resultados foram inferiores na métrica ROUGE ao treinarmos o modelo localmente com a mesma base, devido, pressupomos, à diferença de hardware. Ainda assim, obteve valores maiores na métrica em três *scores* se comparados a outros estados da arte da época como a proposta por [3].

Tabela II – Resultados obtidos na base de dados arXiv.

Modelo	RG-1	RG-2	RG-3	RG-L
Original	35.80	11.05	3.62	31.80
Local	29.18	7.35	2.6	26.38

5. Conclusão Parcial

O trabalho de [4] apresenta uma metodologia de rede neural recorrente para resumos abstratos de artigos científicos. Resultados preliminares obtidos por nós, na mesma base, mostram que a rede proposta gera resumos parecidos com a de um humano e mesmo localmente obteve valores maiores na métrica ROUGE em três *scores* se comparados a outros estados da arte da época. As etapas futuras consistem em finalizar a base de dados com pelo menos 1000 artigos e treinar a base para resumir a metodologia. Para critérios de comparação, o modelo fornecido por [4] será atualizado para os softwares da atualidade (Keras, por exemplo) e compará-lo com o modelo original.

6. Referências

[1] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. ArXiv, abs/1509.00685, 2015.

[2] Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summariza-tion with attentive recurrent neural networks. InProceedings of the 2016 Conference ofthe North American Chapter of the Association for Computational Linguistics: HumanLanguage Technologies, pages 93–98, June 2016.

[3] Ramesh Nallapati, Bowen Zhou, et. al. Abstractive text summarization using sequence-to-sequence RNNs and beyond. InProceedings of The 20th SIGNLL Conference on Computational Natural Language Le-arning, pages 280–290, August 2016.

[4] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, et. al..A discourse-aware attention model for abstractive summari-zation of long documents. In NAACL-HLT, 2018.

Agradecimentos

Ao Centro Universitário FEI pelo empréstimo de equipamentos

[1] Aluno da Maratona de Programação.