# Correção Automática de Acrônimos Sem Explicação

**Daniel Augusto das Neves Carrijo[1], Márcio de Souza Dias[1]**

[1]Departamento de Ciências da Computação - Universidade Federal de Goiás (UFG)
Catalão – GO – Brazil

daniel.carrijo@outlook.com, marciosouzadias@gmail.com

***Abstract.*** *The multi-document summaries may be generated with some linguistic problems, mainly because the automatic summarizers do not have an efficient module that solves such problems. In view of this fact, this work proposes the development of a prototype that corrects one of the most frequent errors in multi-document summaries, the Acronym Without Explanation. Using a multi-document summary corpus, the prototype has obtained an accuracy of 93.5 % in error correction*

***Resumo.*** *Os sumários multidocumento podem ser gerados com alguns problemas linguísticos, principalmente porque os sumarizadores automáticos não contam com um módulo eficiente que solucionam tais problemas. Diante desse fato, esse trabalho propõe o desenvolvimento de um protótipo que corrija um dos erros de maior frequência em sumários multidocumento, o Acrônimo Sem Explicação. Utilizando um córpus de sumários multidocumento, o protótipo obteve uma acurácia de 93,5% na correção do erro.*

## 1. Introdução

Historicamente, a escrita é uma das formas mais usadas na comunicação, sendo utilizada em artigos, mensagens, livros e muitos outros. Atualmente, muitos veículos de informação utilizam-se da escrita para passar uma notícia, tanto em jornais, como em revistas, portais online, e outros. Desta forma, a Sumarização Automática Multidocumentos (SAM) vem tendo um papel muito importante na captação otimizada de informações.

A SAM consiste em extrair as informações chaves de cada texto que tratam do mesmo assunto e agrupá-las da melhor maneira possível em apenas um texto [Mani 2001], de forma que o leitor tenha todas as informações importantes em apenas um texto. Entretanto, a SAM não é uma tarefa fácil e muitos problemas ainda estão presentes, como a produção de um texto totalmente coerente e compreensível.

Estudos como [Koch 1998],[Koch and Travaglia 2002],[Otterbacher et al. 2002], [Pitler et al. 2010], [Kaspersson et al. 2012], [Friedrich et al. 2014] e [Dias 2016] listaram problemas linguísticos que podem ocorrer em textos gerados de forma automática (sumarização automática, sistemas de perguntas/respostas, etc), e que podem prejudicar a qualidade textual.

Segundo Dias (2016), o erro linguístico de maior frequência presente em sumários do Português do Brasil gerados pela SAM é o Acrônimo Sem Explicação. Esse erro consiste em citar um acrônimo (sigla) sem que a explicação venha posteriormente ou anteriormente ao mesmo. Na Figura 1 é mostrado um sumário do córpus CSTNews

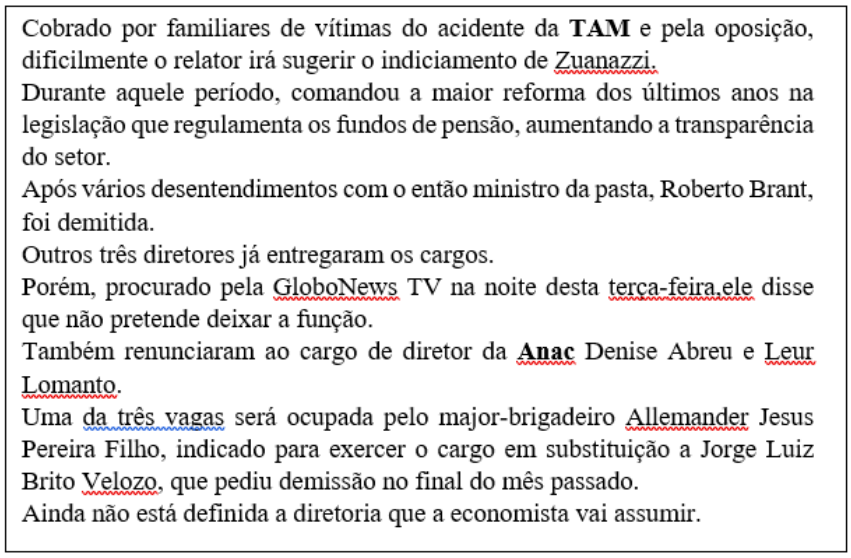
Cobrado por familiares de vítimas do acidente da **TAM** e pela oposição, dificilmente o relator irá sugerir o indiciamento de Zuanazzi.
Durante aquele período, comandou a maior reforma dos últimos anos na legislação que regulamenta os fundos de pensão, aumentando a transparência do setor.
Após vários desentendimentos com o então ministro da pasta, Roberto Brant, foi demitida.
Outros três diretores já entregaram os cargos.
Porém, procurado pela GloboNews TV na noite desta terça-feira,ele disse que não pretende deixar a função.
Também renunciaram ao cargo de diretor da **Anac** Denise Abreu e Leur Lomanto.
Uma da três vagas será ocupada pelo major-brigadeiro Allemander Jesus Pereira Filho, indicado para exercer o cargo em substituição a Jorge Luiz Brito Velozo, que pediu demissão no final do mês passado.
Ainda não está definida a diretoria que a economista vai assumir.

**Figure 1. Sumário automático multidocumento com Acrônimo Sem Explicação**

[Cardoso et al. 2011] em que os acrônimos "TAM" e "Anac"(em negrito) não têm suas explicações (significado) explicitadas no sumário.

Para o erro de Acrônimo Sem Explicação, o qual foi anotado manualmente em um córpus de sumários gerados por sumarizadores automáticos multidocumento [Dias 2016], não foi encontrado na literatura, até o momento, trabalhos que revisam esse tipo de erro de maneira automática. Desta forma, neste artigo nós propomos a criação de um protótipo que automatiza a tarefa de explicitar o significado de um acrônimo, cuja a sua explicação não está presente no sumário. Tal abordagem pode ser útil em geradores automáticos de texto (sumarizadores, sistema de perguntas e respostas, etc), uma vez que tais sistemas não contam com módulos de tal natureza na geração de textos.

Este artigo está organizado da seguinte maneira: na Seção 2 são descritos os trabalhos relacionados; a Seção 3 apresenta o córpus utilizado; a Seção 4 apresenta a metodologia de desenvolvimento; na Seção 5 são discutidos os experimentos e resultados; na Seção 6 é apresentada uma breve conclusão.

## 2. Trabalhos Relacionados

Até o momento, não foi encontrado na literatura trabalhos que busquem sugerir uma solução automática para erros linguísticos do tipo do Acrônimo Sem Explicação, este artigo mencionará brevemente os trabalhos de referência na identificação manual de erros linguísticos que prejudicam a coerência dos textos gerados automaticamente.

Otterbacher et al. (2002) estudaram os problemas relacionados a coesão textual em textos extraídos de sumários multidocumento, e propôs soluções para melhorá-la.

Kaspersson et al. (2012) investigaram os erros que ocorrem em sumários que são oriundos de um documento único, porém o foco foi em expressões de referência que não foram referenciadas e também investigaram como as partes textuais nos sumários são conectadas e, além disso, como o tamanho de um sumário pode interferir na ocorrência

de cada tipo de erro.

Friedrich et al. (2014) apresentaram um córpus de sumários multidocumento, chamado LQVSumm. E ele tratou basicamente de dois erros, o primeiro foi de menção de entidades (que é relacionado à problemas de referência) e o outro que envolve erros de gramática e redundância.

Dias (2016) desenvolveu um classificador automático de coerência textual para sumários multidocumento para o Português do Brasil. Além do mais, o autor fez um estudo sobre os erros que afetam a coerência dos sumários multidocumento gerados por sumarizadores automáticos.

Os trabalhos apresentados nessa seção apenas identificaram manualmente erros que afetam a qualidade do texto. Neste trabalho, o erro de Acrônimo Sem Explicação foi erro escolhido, devido a sua frequência apresentada nos trabalhos relacionados, para um tratamento automático na sua correção.

## 3. Córpus

Este trabalho utilizou o córpus CSTNews [Cardoso et al. 2011]. Córpus foi criado para trabalhar com a sumarização multidocumento. Foram utilizados 4 sumarizadores automáticos para gerar sumários para cada um dos 50 conjuntos (*clusters*) do CSTNews. Uma vez que cada conjunto possui de 2 à 3 textos, totalizando 140 textos fontes, sendo que texto fonte tem uma média de 334 palavras.

Os textos que compõe o córpus são basicamente textos jornalísticos obtidos das páginas web dos maiores jornais do país, como “O Globo”,“Jornal do Brasil”, “Estadão” etc, sendo que os 50 conjuntos reúne textos dos mais variados temas, como política, ciência, esporte, etc. De acordo com os autores, essas fontes foram escolhidas por terem notícias atuais e devido às suas popularidades.

No processo de criação, participaram especialistas da área de Linguística e da Ciência da Computação que fizeram a anotação manual do córpus de diversas informações linguísticas, inclusive a de erros que afetam a qualidade linguística dos sumários automáticos multidocumento.

Para a anotação de erros linguísticos foram utilizados 200 sumários gerados automaticamente, já que para cada *cluster* foram criados 4 sumários diferentes, um de cada sumarizador automático (GistSumm [Filho et al. 2007], RSumm [Ribaldo 2013], RC-4 [Cardoso et al. 2015] e MTRST-MCAD [Castro Jorge 2015]). Na Tabela 1 é mostrado os dados do córpus de sumários automáticos.

**Table 1. Dados do córpus de sumários automáticos multidocumento**

| Sistema | Média de palavras | Média de senteças |
|---|---|---|
| GistSumm | 362 | 11 |
| RSumm | 134 | 4 |
| RC-4 | 132 | 4 |
| MTRST-MCAD | 139.78 | 7.92 |

## 4. Metodologia de Desenvolvimento

Paralelamente à este estudo, um trabalho de identificação do erro Acrônimo Sem Explicação também estava sendo desenvolvido. Assim, os acrônimos que não continham a sua explicação no sumário eram identificados automaticamente. Dessa forma, uma lista foi gerada contendo tais acrônimos.

De posse da lista de Acrônimos Sem Explicação, utilizamos a Wikipedia[1] como base de conhecimento para descobrir o significado (explicação) de tais acrônimos. Para isso, a API da Wikipedia da linguagem de programação Python[2] foi utilizada para trazer as páginas da Wikipedia que contivessem o acrônimo a ser resolvido. A escolha da Wikipedia deve-se ao fato de que a mesma possui um grande acervo de dados e as outras APIs não possuem a mesma quantidade de contéudo disponível para busca.

Como observado, os acrônimos normalmente aparecem no primeiro parágrafo das páginas resultantes das buscas no Wikipedia. Desta forma, o parágrafo de cada página foi recuperado no intuito de determinar o significado do acrônimo em análise, assim que encontramos o acrônimo em uma página da Wikipedia, verificamos se ele está entre parentêses, vírgulas, travessões. Caso esteja, verificamos a localização da explicação, ou seja, antes ou depois do acrônimo. Caso não esteja, verificamos se entre parênteses, vírgulas, ou entre travessões, há uma possível explicação.

Nas Figuras 2 e 3 estão dois textos oriundos da pesquisa dos acrônimos “Anatel” e “Infraero” (em negritos e suas explicações sublinhadas), respectivamente, exemplificando os dois casos citados anteriormente.

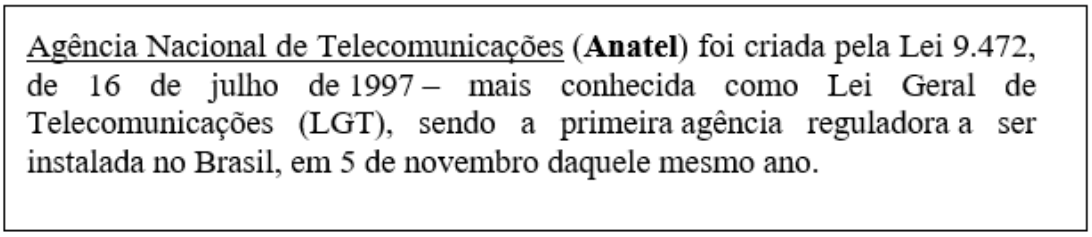
Agência Nacional de Telecomunicações (**Anatel**) foi criada pela Lei 9.472, de 16 de julho de 1997 – mais conhecida como Lei Geral de Telecomunicações (LGT), sendo a primeira agência reguladora a ser instalada no Brasil, em 5 de novembro daquele mesmo ano.

**Figure 2. Primeiro Caso: Acrônimo entre parêteses**

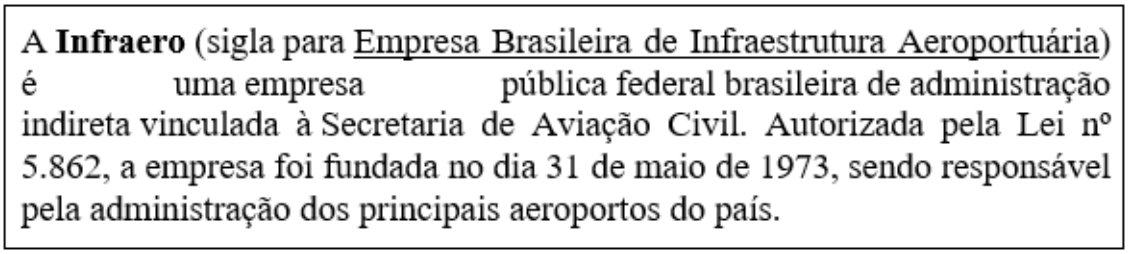
A **Infraero** (sigla para Empresa Brasileira de Infraestrutura Aeroportuária) é uma empresa pública federal brasileira de administração indireta vinculada à Secretaria de Aviação Civil. Autorizada pela Lei nº 5.862, a empresa foi fundada no dia 31 de maio de 1973, sendo responsável pela administração dos principais aeroportos do país.

**Figure 3. Segundo Caso: Explicação entre parêteses**

Em seguida, verifica-se se uma das possiblidades seguintes são verdadeiras: i) se todas as letras do acrônimo são maiúsculas; ii) se pelo menos uma das letras é minúscula.

[1]https://www.wikipedia.org
[2]https://www.python.org

Caso todas as letras do acrônimo sejam maiúsculas, há uma grande probabilidade de que a explicação contenha uma palavra para cada letra do acrônimo. Por exemplo, CBF - Confederação Brasileira de Futebol ou ONU - Organização das Nações Unidas.

Sabendo disso, uma verificação se as letras maiúsculas iniciais da explicação encontrada correspondiam as letras do acrônimo era necessária. Caso todas as letras comparadas fossem as mesmas, o protótipo encontrou a explicação do acrônimo.

Quando a sigla tem letras em minúsculo, a probabilidade de que cada letra corresponde a uma palavra na explicação é quase nula. Por exemplo, o acrônimo "Anatel" possui apenas a primeira letra em maiúscula, e o seu significado encontrado é Agência Nacional de Telecomunicações.

Portanto, ao encontrar alguma menção ou ocorrência do acrônimo em uma página da Wikipedia, apenas utilizamos a explicação encontrada entre parênteses, pois a chance de que a explicação seja a correta é alta, devido ao fato de que o motor de busca da Wikipedia sempre traz as páginas mais pesquisadas no topo da busca. Além disso, devemos respeitar a regra de que a primeira palavra da explicação tem que começar com letra maiúscula.

A seguir, algumas regras foram desenvolvidas no intuito de ampliar a resolução do erro do Acrônimo sem Explicação.

### 4.1. Regra dos Estados

Como o córpus CSTNews é composto por textos jornalísticos, e o mesmo possui vários assuntos, inclusive os relacionados a política, há referências a diversos políticos dos mais variados estados brasileiros. Na Figura 4 é mostrado um trecho de um sumário retirado do CSTNews, onde um acrônimo relacionado a um partido político juntamente com um acrônimo relacionado a um estado brasileiro (em negrito) ocorre.

> "Por unanimidade, os integrantes da Mesa Diretora do Senado decidiram aceitar o pedido do Conselho de Ética de encaminhar para a Polícia Federal pedido para realização de nova perícia nos documentos do presidente da Casa, Renan Calheiros **(PMDB-AL)**."

**Figure 4. Sumário automático multidocumento com Acrônimo Sem Explicação**

Quando o acrônimo sucede um hífen e tem apenas dois caracteres, ele tem grandes chances de ser um estado, portanto a pesquisa na Wikipedia se tornaria desnecessária. Portanto, quando isso ocorre, utilizamos uma lista de estados para verificar se tal acrônimo é realmente um estado, e caso não seja o protótipo volta a utilizar a Wikipedia.

### 4.2. Regra do Título

Algumas vezes, a referência ao acrônimo não está no primeiro parágrafo da página de retorno da busca do Wikipedia, e procurá-lo em outros parágrafos demandariam um alto custo e possivelmente o acrônimo não seria encontrado. Dessa forma,o próprio título da página pode ser a explicação do acrônimo desejado.

Para verificar se o título realmente é a explicação do acrônimo em análise, foi realizado o procedimento de comparar as letras do acrônimo com as primeiras letras de cada palavra da possível explicação. Caso as letras sejam as mesmas, o protótipo irá considerar que a explicação foi encontrada.

## 5. Experimentos e Resultados

Para realizar os experimentos, 92 sumários dos 200 foram utilizados. Essa quantidade é devido a presença do erro de Acrônimo sem Explicação.

Com o objetivo de desenvolver e avaliar o protótipo, 68 sumários (escolhidos aleatoriamente) foram utilizados no treinamento e 24 sumários foram utilizados para teste. Na Tabela 2 é apresentado os resultados obtidos:

**Table 2. Resultado dos experimentos**

| Quantidade de Sumários de Teste | Acrônimos encontrados | Acrônimos explicados |
|---|---|---|
| 24 | 77 | 72 |

Como podemos observar na Tabela 2, obtivemos 93,5% de acurácia. Um resultado que podemos considerar muito bom, uma vez que trabalhamos apenas com heurística e nada muito complexo, como Redes Neurais, Aprendizado de Máquina, etc.

Para certificar de que a explicação trazida pelo protótipo era a correta, nós nos baseamos no contexto do sumário que continha um acrônimo não explicado e confrontamos com a explicação apresentada, e assim, concluímos que em 72 das 77 occorências de um acrônimo sem explicação, a explicação dada pelo protótipo estava correta.

Os únicos três casos de acrônimos que não foram explicados foram: CGE[3](duas ocorrências), NHK[4](uma ocorrência), P-SOL[5], escrito dessa maneira (duas ocorrências).

No primeiro caso, não tem nenhuma ocorrência do acrônimo CGE no Wikipedia, e por isso no momento da busca, não houveram resultados compatíveis com CGE.

O NHK é um acrônimo japonês, e a Wikipedia utilizada neste trabalho está em Português, o que inviabilizou a busca correta da explicação para o acrônimo NHK.

Já no caso do acrônimo P-SOL, ele não é detectado pelo protótipo devido a sua escrita está diferente da usual(PSOL).

## 6. Conclusão

Esse trabalho inova ao propor um protótipo que corrige um erro da qualidade linguística, Acrônimo Sem Explicação, de maneira automática. Com uma acurácia de 93,5%, acreditamos que a abordagem utilizada nesse trabalho obteve sucesso. Em contrapartida, o protótipo está dependente de um recurso sujeito a suas variações de conteúdo e sua possível instabilidade funcional, que é a Wikipedia.

[3]Centro de Gerenciamento de Emergências
[4]Nihon Hikikomori Kyōkai
[5]Partido Socialismo e Liberdade

Para diminuir essa dependência e possivelmente melhorar a acurácia do protótipo, propomos como trabalho futuro o uso de similaridade semântica para contexto textual de forma mais efetiva.

## References


Cardoso, P., Castro Jorge, M., and Pardo, T. (2015). Exploring the rhetorical structure theory for multi-document summarization. In *Proceedings of the 5th Workshop RST and Discourse Studies*, pages 1 – 10.

Cardoso, P., Mazieiro, E., Jorge, M., Seno, E., di Felippo, A., Rino, L., Nunes, M., and Pardo, T. (2011). 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*, pages 88–105.

Castro Jorge, M. L. R. (2015). *Modelagem gerativa para sumarização automática multidocumento*. PhD thesis, Instituto de Ciências Matemáticas e de Computação - ICMC/USP.

Dias, M. S. (2016). *Investigação de modelos de coerência local para sumários multidocumento*. PhD thesis, Instituto de Ciências -USP.

Filho, P. P. B., Pardo, T. A. S., and das Graças Volpe Nunes, M. (2007). Sumarização automática de textos científicos: Estudo de caso com o sistema gistsumm. Technical report, NILC - ICMC-USP. 23 p.

Friedrich, A., Valeeva, M., and Palmer, A. (2014). Lqvsumm: A corpus of linguistic quality violations in multi-document summarization. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kaspersson, T., Smith, C., Danielsson, H., and Jönsson, A. (2012). This also affects the context - errors in extraction based summaries. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Koch, I. G. V. (1998). *A coesão textual – Mecanismos de Constituição Textual, A organização do Texto, Fenômenos de Linguagem*. Linguística Contexto – Repensando a Língua Portuguesa, 10 edition.

Koch, I. G. V. and Travaglia, L. C. (2002). *A coerência textual*. Editora Contexto.

Mani, I. (2001). *Automatic summarization*, volume 3. John Benjamins Publishing.

Otterbacher, J. C., Radev, D. R., and Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: A preliminary study. In *Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4*, AS '02, pages 27–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pitler, E., Louis, A., and Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the*

*Association for Computational Linguistics*, ACL '10, pages 544–554, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ribaldo, R. (2013). Investigação de mapas de relacionamento para sumarização multi-documento. Monografia de Conclusão de Curso, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Novembro, 61p.