AvaliaGeo: Sistema para Validação de Topônimos em Notícias

Arthur Felipe de Freitas Domingues[1], **Bruno Rabello Monteiro**[1]

[1]Instituto de Ciências Exatas e Aplicadas – Universidade Federal de Ouro Preto(UFOP)

arthurfreitas772@gmail.com, bruno@ufop.edu.br

Abstract. *Solutions to extracting geographic information problems from texts and documents often need labeled databases to carry out experiments or validate algorithms. However, many of these databases are not costless made available for use. This work aims to facilitate the generation of geographically labeled databases, using voluntary contributions to disambiguate toponyms present in the news. We propose using the Cronbach's Alpha coefficient to validate the contributions, considering each news item a questionnaire and each toponym candidate a questionnaire item. Preliminary experiments achieved 70% reliability in the disambiguation of toponyms for database generation.*

Resumo. *Soluções para problemas de extração de informação geográfica de textos e documentos precisam, muitas vezes, de bases de dados rotuladas para realização de experimentos ou para validação de algoritmos. Entretanto, muitas dessas bases não são gratuitas ou não são deixadas disponíveis. Este trabalho tem por objetivo facilitar a geração de bases de dados rotuladas geograficamente, com o uso de contribuições voluntárias para a desambiguação dos topônimos presentes nas notícias. Para validar as contribuições é proposto o uso do coeficiente Alfa de Cronbach, considerando cada notícia um questionário e cada candidato à topônimo um item desse questionário. Experimentos preliminares alcançaram 70% de confiabilidade na desambiguação dos topônimos para geração de bases de dados.*

1. Introdução

A utilização de informação geográfica em aplicações vem aumentando, impulsionada pela possibilidade de desenvolvimento de sistemas personalizados [Larsen 2010], ou mesmo pela popularização de serviços como o Google Maps[1], Waze[2] e outras ferramentas de navegação. Soluções para o Problema de Resolução de Escopo Geográfico (PREG) correspondem a uma das áreas que fazem uso desses dados geográficos presentes em textos e documentos. A resolução do PREG visa determinar o escopo geográfico dos documentos, utilizando os topônimos (nomes de lugares) ou qualquer outra forma de informação geográfica contida no texto.

As tarefas *Geoparsing*, *Resolução de Referência* e *Fixação das Referências* são necessárias para realizar a determinação do escopo geográfico [Monteiro et al. 2016]. O *Geoparsing* objetiva identificar referências à lugares, como os topônimos. Comumente, é

[1]https://www.google.com.br/maps

[2]https://www.waze.com/

necessário desambiguar essas referências geográficas encontradas, seja porque diferentes locais compartilham um mesmo topônimo [3], seja porque nomes comuns e próprios são usados para dar nomes aos locais [4]. A desambiguação corresponde à tarefa *Resolução de Referência*. A última tarefa é a *Fixação das Referências* que constrói propriamente o escopo geográfico(um par com as coordenadas do lugar mais importante, ou a região contendo todos os topônimos encontrados, por exemplo).

Um dos problemas no desenvolvimento de soluções para o PREG é a falta de bases de dados que possam ser usadas gratuitamente nos experimentos [Gritta et al. 2018]. O uso de bases de dados proprietárias, não compartilhadas, dificulta a comparação de trabalhos e soluções [Monteiro et al. 2016] e [Gritta et al. 2018]. Neste sentido, o AvaliaGeo é um sistema desenvolvido para permitir a contribuição voluntária na tarefa de *Resolução de Referências*, objetivando gerar bases de dados textuais rotuladas geograficamente. A validação das contribuições é feita utilizando-se o coeficiente Alfa de Cronbach [Cronbach 1951]. De um modo resumido, esse coeficiente é usado para verificar como itens estão correlacionados, especificamente para este trabalho, esse coeficiente mede o quão correlatas estão as contribuições voluntárias.

2. Uso do coeficiente Alfa de Cronbach

Vários trabalhos utilizam o coeficiente Alfa de Cronbach para medir a consistência das respostas obtidas com questionários. Historicamente, a maior utilização desse coeficiente está em trabalhos das áreas de saúde como a Medicina, a Psicologia, e a Enfermagem, seguido por trabalhos nas áreas de Ciências Sociais e Economia [Matthiensen 2010].

Por exemplo, em [Almeida et al. 2010], o coeficiente Alfa de Cronbach verifica a confiabilidade das respostas dada aos questionários sobre os níveis de satisfação nas Unidades Básicas de Saúde, feitas pelos próprios funcionários da rede de saúde pública. Já em [Freitas and Rodrigues 2005], o coeficiente é usado para validar a avaliação institucional feita pelo corpo docente da Universidade Estadual do Norte Fluminense (UENF).

Este trabalho adapta o uso do coeficiente Alfa de Cronbach para validar as contribuições voluntárias na desambiguação de topônimos em textos, a fim de gerar bases de dados rotuladas geograficamente.

3. Desenvolvimento

O AvaliaGeo[5] utiliza a arquitetura cliente/servidor, e foi construído usando HTML5, CSS3 e JavaScript. O servidor foi projetado com o framework Flask[6] junto da linguagem Python. A tarefa de *Geoparsing* sobre as notícias é de responsabilidade da biblioteca Polyglot[7]. O armazenamento das contribuições voluntárias é feito com o Firebase Storage. O sistema está hospedado na plataforma Heroku[8].

O AvaliaGeo é um sistema semi-automático que necessita de três etapas para sua execução: Pré Processamento (1) , Coleta de Contribuições (2), e Pós Processamento (3).

[3]Por exemplo: ”Paris”é tanto a capital da França quanto uma cidade do estado do Texas, EUA

[4]Por exemplo: ”Mariana”e ”Esmeraldas”são cidades no estado de Minas Gerais

[5]Código-fonte disponível em: https://github.com/artcomp/Avaliageo

[6]https://flask.palletsprojects.com/

[7]https://pypi.org/project/polyglot/

[8]http://avaliageo.herokuapp.com/

Na etapa (1) é necessário obter as notícias, em português, que irão constituir as bases de dados. O *crawling* é feito fora do AvaliaGeo. Ainda na etapa (1), o AvaliaGeo reconhece os candidatos à topônimos presentes nas notícias e utiliza uma fonte de conhecimento externa, contendo os dados geográficos para desambiguação como, por exemplo, o *gazetteer* GeoNames[9](utilizado nesse trabalho) ou o *Open Street Map*[10].

Durante a segunda etapa (2), as contribuições voluntárias são coletadas. Para uso do coeficiente Alfa de Cronbach, cada notícia é tratada como um questionário, e cada candidato à topônimo da notícia é considerado como uma pergunta (item) do questionário. Os voluntários resolvem a tarefa de *Resolução de Referências* selecionando, em cada notícia, a referência geográfica que desambigua cada um dos candidatos. Além disso, o voluntário informa o grau de certeza da sua desambiguação através de uma escala discreta-ordinal (0%,25%,50%,75%,100%).

Na parte (3), o AvaliaGeo mapeia as desambiguações feitas por cada voluntário (contribuições) em números para o cálculo do coeficiente o Alfa de Cronbach. Para isso, é criada uma matriz X, $n \times k$, para cada notícia (questionário), em que: n indica a quantidade de respostas do questionário (número de topônimos desambiguados), e k é o número de itens do questionário (quantidade de alternativas para desambiguar cada candidato à topônimo). O cálculo do Alfa de Cronbach usou a Equação 1. Na equação, σ_i^2 é a variância de cada coluna da matriz X; σ_τ^2 é a variância da soma de cada linha de X.

O valor mínimo considerado aceitável para o Alfa é 0,70. Em contrapartida, o valor máximo esperado é 0,90. Acima deste valor, pode-se considerar que há redundância ou duplicação. Valores de Alfa de Cronbach entre 0,80 e 0,90 são os desejados, segundo [Streiner 2003].

$$\alpha = \frac{k}{k-1}\left[\frac{\sigma_\tau^2 - \sum_{i=1}^{k}\sigma_i^2}{\sigma_\tau^2}\right] \quad (1)$$

Ao final da etapa (3), obtém-se a solução para cada candidato à topônimo e a média da confiabilidade das contribuições. Para determinar a resposta correta utiliza-se a maioria simples dentre as contribuições. O número de contribuições (voluntários) é quantificado para cada notícia. Determinou-se, via método de inspeção, que cada notícia deveria ter um mínimo de 10 contribuições distintas [11].

Para uma notícia ser aceita, ou seja, considerada rotulada e desambiguada, é necessário que o valor do coeficiente Alfa de Cronbach seja superior à 0,7 e a escala de confiabilidade superior à 50%. As notícias aceitas são armazenadas, de modo a compor uma base de dados, que poderá, futuramente, ser disponibilizada.

4. Resultados

Experimentalmente, utilizou-se um dataset de vinte (20) notícias em português do portal G1[12], no editorial Brasil. O AvaliaGEo ficou disponível para receber contribuições por 45

[9]https://www.geonames.org/

[10]https://www.openstreetmap.org/

[11]Não existe um consenso na literatura a respeito de um valor mínimo obrigatório.

[12]https://g1.globo.com/

dias, e recebeu 228 contribuições. Ao final, 14 das 20 notícias foram consideradas aceitas (totalmente desambiguadas). Na média, cada notícia teve mais de 11 contribuições. Uma verificação manual foi feita sobre as notícias e constatou que as 14 notícias foram desambiguadas corretamente.

A validação, com uso do coeficiente Alfa de Cronbach, não determina exatamente se a solução para desambiguação está correta, mas sim determina se as contribuições voluntárias convergem. Dessa forma, este trabalho propõe a utilização dos conhecimentos e experiências prévias dos voluntários como ferramenta de validação, já que este método permite uma maior velocidade e escalabilidade na geração de bases rotuladas geograficamente.

Foi possível notar que, em alguns casos, os coeficientes Alfa de Cronbach apresentam valores superiores à 0.9. Entretanto, esses valores são esperados, já que, na própria literatura, são justificados por itens duplicados no questionário. Neste caso, a duplicação significa um mesmo topônimo aparecendo mais de uma vez na notícia. Assim, a presença de topônimos repetidos, aliados com determinada consistência na resposta dos usuários, podem levar a valores superiores a 0.9. Entretanto, esses valores são aceitáveis.

Como trabalhos futuros, pretende-se aumentar a experimentação: com o uso de um número maior de notícias; com a possibilidade de contribuições voluntárias para melhoria da tarefa de *Geoparsing*(com o informe de topônimos não encontrados); com a possibilidade de carregamento de bases de dados para serem validadas pelos usuários e com a utilização de outras soluções para a identificação dos topônimos. Por fim, evoluir o AvaliaGeo, permitindo uma execução totalmente automática das etapas.

Referências

Almeida, D., Santos, M. d., and Costa, A. F. B. (2010). Aplicação do coeficiente alfa de cronbach nos resultados de um questionário para avaliação de desempenho da saúde pública. *XXX Encontro Nacional de Engenharia de Produção*, 15:1–12.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Freitas, A. L. P. and Rodrigues, S. G. (2005). A avaliação da confiabilidade de questionários: uma análise utilizando o coeficiente alfa de cronbach. In *Anais... XVII SIMPEP,*.

Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623.

Larsen, N. (2010). Market segmentation - a framework for determining the right target customers. Bachelor's thesis, Aarhus School of Business, Aarhus BSS, Denmark.

Matthiensen, A. (2010). Uso do coeficiente alfa de cronbach em avaliações por questionários. *Embrapa Roraima-Documentos (INFOTECA-E)*.

Monteiro, B. R., Jr., C. A. D., and Fonseca, F. T. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96:23–34.

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*, 80(1):99–103.