Uma Aplicação de Sumarização Textual Automática para Dispositivos Móveis

Luciano Cabral, Manoel Neto, Henrique Nunes, Israel Junior e Marlysson Oliveira

Instituto Federal de Pernambuco (IFPE) – *Campus* Jaboatão dos Guararapes
Av. Barão de Lucena, 251 - Centro, Jaboatão dos Guararapes - PE, 54110-005

{lscabral, henriquenunesti}@gmail.com@gmail.com, {manoel-neto-, israelljunnior15987426, marlysson.oliveira}@hotmail.com

Abstract. *This paper describes the development of an application for content summarization for mobile devices, aiming to help society in the search, understanding and interpretation of texts whose core is not clear enough. Generating summaries is an activity that creates a small portion of a given text, being small but sufficient to cover the most important parts of the text. Among the applications, Portuguese language studies are listed, as well as assisting in the process of learning a second foreign language, demonstrating the textual nucleus in different languages.*

Resumo. *Este artigo descreve o desenvolvimento de uma aplicação para sumarização de conteúdos para dispositivos móveis, objetivando auxiliar a sociedade na busca, compreensão e interpretação de textos cujo seu núcleo não esteja suficientemente claro. A geração de sumários é uma atividade que cria uma pequena porção de um determinado texto, sendo pequeno, mas suficiente para cobrir as partes mais importantes do texto. Dentre as aplicações, enumeram-se estudos de língua portuguesa, além de auxiliar no processo de aprendizagem de uma segunda língua estrangeira, demonstrando o núcleo textual em diferentes idiomas.*

1. Introdução

A rápida disseminação da internet produziu uma enorme quantidade de informações disponíveis, principalmente quando se trata de documentos textuais, e.g. notícias, livros, sinopse de filmes, dentre outros. Este grande volume pode se tornar um problema devido à qualidade do conteúdo apresentado. Desta forma surgiu a necessidade da utilização de métodos automáticos, para entender, classificar e apresentar de forma clara e concisa as informações consideradas mais relevantes. A sumarização automática de textos tornou-se uma possível ferramenta para solucionar este problema.

A Sumarização de textos (ST) é o processo de criação automática de uma pequena versão de um ou mais documentos de texto (FERREIRA, et al., 2013). A capacidade de sumarizar automaticamente um conteúdo é um trabalho complexo de mineração de texto. Pesquisas na área iniciaram-se em 1958 com Luhn (LUHN, 1958), que propôs analisar frequências e distribuições de palavras para calcular a importância das sentenças para criação de resumos. A necessidade cada vez maior de sumarização automática de documentos cativou mais e mais pesquisadores para a área (NENKOVA & MCKEOWN, 2011) (ABUOBIEDA *et al.*, 2013).

A sumarização automática ainda continua sendo foco de grandes pesquisas, sendo elas, geralmente, relacionadas às máquinas com grande poder de processamento, e.g. notebooks e desktops. Desta maneira, é possível utilizar várias ferramentas para a obtenção de uma maior precisão e qualidade, e.g. *Stanford Parser* ou *GATE*. Entretanto, o atual estado da arte carece de sumarizadores eficientes para dispositivos móveis, sendo os smartphones os mais usuais.

Visando os dispositivos que são equipados com um hardware menos potente, esta pesquisa concentra-se na produção de um software de sumarização automática para *smartphones* que utilizam o sistema operacional *Android, Windows Phone* e *iOS*. O escopo deste trabalho apresenta os resultados alcançados no desenvolvimento do protótipo para dispositivos *Android*.

2. Materiais e Métodos

A pesquisa foi iniciada buscando materiais necessários para a compreensão de técnicas e métodos necessários para desenvolvimento do protótipo, para isso, buscou-se entender o estado da arte atual da área. Após realização dos estudos, criou-se e testou-se várias hipóteses acerca do assunto estudado. Ao obter-se os resultados desejados começa-se o desenvolvimento do protótipo com os melhores métodos encontrados. Ao todo, foram utilizados sete métodos multilíngues de sumarização automática de texto, que foram escolhidos após os testes e estudos realizados, e toma-se como base a eficiência de cada um. Além de ser uma plataforma de sumarização, conseguiu-se inserir no *protótipo* um modo de aprendizado de língua estrangeira, onde você pode usar o texto original e a tradução para melhor compreensão de outros idiomas.

3. Arquitetura funcional do protótipo

A arquitetura funcional do protótipo proposto é estruturada em cinco componentes: Interface gráfica do usuário, pré-processamento, combinação e ranqueamento de métodos, pontuação e tradução.

- *Serviços de Pré-processamento*: Esse módulo fornece serviços de processamento de linguagem natural de texto para o pré-processamento de texto, incluindo separação de palavras e sentenças, e remoção de ruídos textuais (e.g. tags HTML e/ou XML) implementando a utilização de expressões regulares (regex), além disso, este serviço conta com um trecho de código de inteligência artificial para definição do idioma do texto de entrada.
- *Pontuação*: Esse serviço é responsável por calcular uma pontuação para cada elemento textual;
- *Métodos de combinações e ranqueamento*: Nesta etapa, é realizado a combinação entre as pontuações produzidas pelo serviço anterior e classifica as frases de acordo com suas pontuações. Além disso, o sistema utiliza a pontuação para eliminar as sentenças com as pontuações mais baixas.
- *Tradução*: Esse passo é realizado utilizando o Google Translate API para fornecer o resumo em um idioma diferente do que a língua original do texto de entrada.

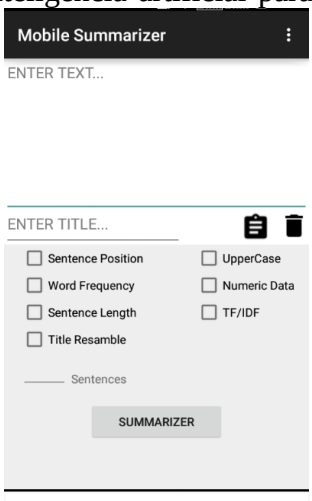

Fig. 1. Tela principal

- *Interface Gráfica*: Ela fornece a interface entre o usuário e o aplicativo. O objetivo principal da interface gráfica é receber, dentre a lista, a notícia à ser sumarizada.

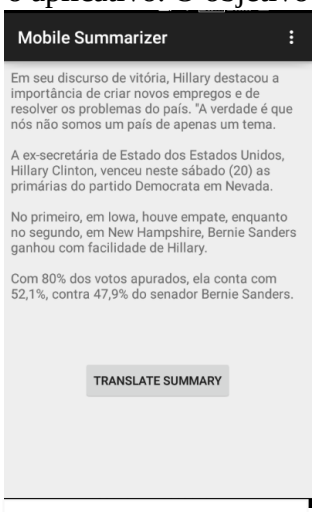

Fig. 2. Sumário gerado

Todos estes componentes contribuem para a nossa hipótese de trabalho em que uma aplicação de sumarização móvel deve ser independente de linguagem e eficiente.

4. Resultados e Discussão

Em (LEITE, 2010) foram avaliados sete sistemas de sumarização da Universidade de São Paulo (USP), utilizando corpus TeMário 2003 – Português (PARDO & RINO, 2003). Entre os sete, o superior, *Supor*, alcançou precisão em termos de *F-Measure* de aproximadamente 0,43.

Buscando e avaliando outros sistemas mobile publicado no *Google Play* encontramos outros aplicativos, como o *Squash – Text Summarizer*, um sumarizador de páginas web, notícias e artigos financeiros. Os resumos gerados pelo *Squash* podem variar de 4 até 8 sentenças. A comparação foi realizada entre sistemas citados anteriormente e o protótipo proposto usando o corpus *TeMário*. Os resultados obtidos são mostrados na tabela 1. Os dados foram obtidos através do ROUGE (Lin, 2004) um software de avaliação de qualidade de sumários textuais.

Tabela 1. ***Comparação entre os sistemas usando o corpus TeMário***

Aplicativo	Precisão	Cobertura	F-Measure
SuPor	-	-	0,43
Squash	0,40	**0,59**	0,45
Protótipo proposto	**0,42**	0,55	**0,48**

Embora estatisticamente quase idêntico, os resultados na Tabela 1 mostram uma ligeira vantagem para o sistema proposto, em termos de precisão e pontuações *F-measure*. Outro experimento foi realizado usando o corpus DUC-2002. Nesse experimento, o DUC *BaseLine* foi utilizado como objetivo de avaliar os sistemas que utilizam esse conjunto de dados em inglês. Os resultados são visíveis na tabela 2.

Tabela 2. ***Comparação entre os o Protótipo proposto e Baseline***

Aplicativo	Precisão	Cobertura	F-Measure
Baseline	0.48	0.48	0.48
Protótipo proposto	0.39	0.58	0.47

Os experimentos em inglês mostram resultados inferiores na precisão em relação aos resultados em português, no entanto o *F-Measure* segue no mesmo nível. Comparando-o ao *Baseline*, os resultados foram consideráveis.

Ao decorrer dos 12 meses de projeto, obtivemos dois artigos publicados em congressos internacionais, no México e na Áustria, o que respalda o trabalho através de um júri qualificado e internacionalmente reconhecido.

- DOI: http://dx.doi.org/10.1109/MICAI.2015.8

- DOI: http://dx.doi.org/10.1145/2960811.2967156

5. Conclusões

Por se tratar de um sistema que foi desenvolvido para dispositivos que dispõe de um hardware com poder de processamento inferior as demais máquinas como desktops e notebooks, o Protótipo proposto possui limitações quando se trata de ferramentas auxiliares para o desenvolvimento dos algoritmos dos métodos de sumarização. Porém os resultados obtidos parcialmente ao longo da concretização do primeiro protótipo foram satisfatórios. Pesquisamos, estudamos, selecionamos e desenvolvemos métodos de sumarização automática para construção do protótipo.

Portanto, após a realização de vários testes podemos concluir que o Protótipo proposto mostrou-se eficiente, e possui grande potencial para competir com aplicativos oriundos de projetos de grandes universidades. O Protótipo proposto encontra-se no Google Play disponível para download, gratuitamente, ele pode ser encontrado pelo nome *Mobile Summarizer*. Em setembro de 2017, o protótipo reunia mais de 1100 downloads, onde apenas 113 destes eram oriundos do país de origem (Brasil), a diferença estava distribuída entre 25 países do mundo. O projeto foi finalizado com êxito e alcançando resultado além do esperado. Em projetos futuros pode ser estudado, avaliado e implementado outros métodos de sumarização automática.

Referências

Abuobieda, A. et al. Opposition Differencial Evolution Based Method for Text Summarization. In Intelligent Information and Database Systems: 5th Asian Conference, ACIIDS 2013, Kuala Lumpur, Malaysia, March 18-20, 2013, Proceedings, Part I, Organizers: Selamat, A.; Nguyen, N.T.; Haron, H. Springer Berlin Heidelberg, 2013.

Ferreira, R., Cabral, L. S., Lins, R. D., Silva, G. F. P., Freitas, F., Cavalcanti, G. D. C., Lima, R. Simske, S. J. and Favaro, L. Assessing Sentence Scoring Techniques for Extractive Text Summarization. Expert Systems with Applications, v. 40, pp. 5755-5764, 2013.

Leite, D. S. Um estudo comparativo de modelos baseados em estatísticas textuais, grafos, e aprendizado de máquina para sumarização automática de textos em português. MSc Dissertation, CCEN-PPCC-UFSCAR. December, 2010.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Proceedings of the workshop on text summarization branches out (was 2004) (pp. 25–26). Barcelona, Spain.

Luhn, H. P. (1958). A business intelligence system. IBM J. Res. Dev. 2, 4 (October 1958), 314-319.

Nenkova, A., & McKeown, K. (2011). Automatic summarization (pp. 103–233). Now Publishers Inc.

Pardo, T. A. S. and Rino, L. H. M. TeMário: Um Corpus para Sumarização Automática de Textos. USP, UFSCAR, UNESP. Technical Report NILC-TR-03-09, October 2003.