UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

Sistema de Sumarização Automática de Textos Baseado em Classes de Documentos

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno: Ihago Henrique Lucena e Silva (ihls@cin.ufpe.br)
Orientador: Prof. Dr. Rafael Dueire Lins (rdl@cin.ufpe.br)
Área: Processamento de Linguagem Natural

Recife, 17 de Abril de 2017

Resumo

A sumarização automática de textos é o processo computacional da geração uma versão mais curta de um ou mais documentos de texto, mantendo a informação fundamental dos documentos originais. Esta área tem sido apontada como uma forma importante de encontrar informações relevantes em grandes bibliotecas de texto, principalmente na Internet.

O trabalho aqui proposto visa encontrar as melhores combinações de métodos de sumarização extrativa, dentre as técnicas mais amplamente difundidas para sumarização, para diferentes classes de documentos.

Introdução

A enorme quantidade de informações disponíveis hoje em grandes bibliotecas digitais e na Internet tornou humanamente impossível selecionar, de forma eficiente, informações úteis. A demanda por ferramentas automáticas capazes de sintetizar informações de forma clara e concisa de documentos textuais passou a ser cada vez maior. A utilização de técnicas de sumarização automática de textos [1] tem sido apontada como uma possível solução para este crescente problema.

O uso de sumarização automática também tem sido estudado como oferecendo potencial melhora nas tarefas de recuperação de informação e classificação de textos, uma vez que reduz o tamanho do documento e, consequentemente, informações possivelmente ruidosas para essas tarefas. A utilização de resumos dos textos, também, diminui consideravelmente o tempo de classificação dos textos [2].

Com o trabalho proposto, espera-se associar a categoria de classificação de um documento como um bom critério para escolher quais técnicas de sumarização extrativa são mais indicadas para cada texto, gerando resumos de melhor qualidade de informação e legibilidade para cada classe de documentos.

Objetivos

O objetivo geral do trabalho proposto é criar um sistema, que a partir de um corpus de textos já classificados em certas categorias e de um conjunto de técnicas de sumarização extrativa, verifique quais as combinações de técnicas produzem os melhores resumos para cada categoria de documento.

Os objetivos específicos desse trabalho incluem:

- Selecionar um corpus de textos já classificados coerentemente, ou seja, um corpus que não apresente classificações muito genéricas;
- Selecionar técnicas de sumarização extrativa de textos;
- Adotar medidas para equilatar e comparar a qualidade dos resumos de um texto produzidos automaticamente;
- Avaliar manualmente e de forma qualitativa uma parte dos resumos produzidos para cada uma das classes;
- Tentar identificar e sistematizar, a partir dos resultados obtidos, as características de cada categoria de texto que justifiquem a combinação selecionada pelo sistema.

Metodologia

A metodologia aqui descrita consiste na definição do corpus de textos, bem como nas metodologias de avaliação dos resultados.

O corpus de texto a ser utilizado será o corpus CNN [3], desenvolvido pelo supervisor deste trabalho e seus colaboradores em no projeto de P&D chamado FLIP, em parceria com a Hewlett-Packard a partir de textos do site da CNN (www.cnn.com). O corpus CNN na versão atual consiste de 3.000 textos em inglês atribuídos. Os textos estão originalmente classificados em 11 categorias: "África", "América Latina", "Ásia", "Esportes", "Estados Unidos da América", "Europa", "Negócios", "Oriente Médio", "Tecnologia", "Viagens" e "Mundo". Tal classificação foi considerada inadequada uma vez que a classe "Mundo", por exemplo, é demasiadamente genérica. Uma nova ontologia foi desenvolvida pelo grupo do projeto FLIP baseada nas classificações da Google e Yahoo. Cada um dos 3.000 textos do corpus CNN foi automaticamente reclassificado segundo a nova ontologia desenvolvida em três categorias em ordem decrescente de importância, sendo tal classificação verificada por três especialistas. Esta nova classificação dos textos será utilizada neste trabalho.

Avaliar a qualidade dos resumos produzidos é algo muito importante e ao mesmo tempo muito complexo. A importância da avaliação da qualidade serve para selecionar as melhores técnicas para um determinado contexto. Uma ferramenta muito utilizada para o propósito da avaliação automática de resumos é o ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). Uma primeira avaliação quantitativa será, então, realizada utilizando o ROUGE, uma vez que o mesmo consegue avaliar automaticamente a qualidade de um resumo comparando-o com outros resumos (ideais) criados por seres humanos. Dessa forma será possível eliminar a subjetividade humana do processo de avaliação nessa etapa [4] e comparar numericamente a qualidade de um resumo.

Uma característica especial do corpus CNN é a existência de um "*gold standard*" [3], um sumário extrativo para cada um dos seus 3.000 textos, que foi gerado a partir de uma metodologia desenvolvida no Projeto FLIP, a partir dos *highlights,* um sumário abstrativo de alta qualidade gerado pelos próprios autores dos textos extraídos do site da CNN. Será realizada, ainda, uma avaliação qualitativa, por pessoas, dos resumos produzidos pelo sistema. O número de casamentos entre as frases do resumo gerado com as frases dos *gold standard* são uma medida da qualidade dos resumos a ser considerada neste trabalho.

A referência [5] apresenta uma análise preliminar da relação entre técnicas de sumarização automática e classificação de textos, porém como foi aí concluído, a classificação original do corpus CNN devido a sua extrema generalidade, não permitiu

uma correlação clara entre o conteúdo do documento (sua classificação) e as técnicas mais adequadas para sua sumarização. Apenas nas classes “Negócios” e “Esportes” ficou evidenciada uma correlação da maior adequabilidade de algumas técnicas de sumarização com o tipo de documento.

Logo, espera-se com o presente trabalho, fazendo uso da nova classificação proposta mais abrangente e detalhada, lançar uma nova luz nessa importante área de pesquisa.

Cronograma

Atividade	Março/17	Abril/17	Maio/17	Junho/17	Julho/17
Revisão bibliográfica	X	X			
Estudo dos algoritmos		X	X		
Implementação dos algoritmos			X	X	
Realização dos experimentos			X	X	
Avaliação dos resultados obtidos			X	X	X
Escrita da monografia			X	X	X
Revisão do texto				X	X
Preparação da apresentação				X	X
Entrega e defesa do trabalho					X

Referências

[1] Ferreira, R.; Freitas, F.; Cabral, L. S. C.; Lins, R. D.; Lima, R.; Franca, G.; Simske, S. J.; Favaro, L. (2014). **A Context Based Text Summarization System**. 11th IAPR International Workshop on Document Analysis Systems, 2014, pp. 66-70.

[2] Koulali, R.; Mahmoud, E.-H.; Abdelouafi, M. **Arabic Topic Detection using Automatic Text Summarisation**. Computer Systems and Applications (AICCSA). pp. 1-4, 2013.

[3] Lins, R. D.; Simske, S.J. ; CABRAL, L. S. ; Silva, G. de F. P. e ; Lima, R. ; Mello, R. F. ; Favaro, L. **A Multi-tool Scheme for Summarizing Textual Documents**. In: IADIS International Conference WWW/Internet 2012, p. 409-414. 2012.

[4] Lin, C.-Y. **ROUGE: A Package for Automatic Evaluation of Summaries**. Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004.

[5] Ferreira, R. ; Lins, R. D. ; Cabral, L. ; Freitas, F. ; Simske, S. J. ; Riss, M. **Automatic Document Classification using Summarization Strategies**. Proceedings of the 2015 ACM Symposium on Document Engineering - DocEng '15. p. 69-124.

Possíveis Avaliadores

Os possíveis avaliadores que irão compor a banca examinadora são:

- Prof. Dr. Rafael Dueire Lins (rdl@cin.ufpe.br) - CIn/UFPE (orientador)
- Prof. Dr. Frederico Luiz Gonçalves de Freitas (fred@cin.ufpe.br) - CIn/UFPE (avaliador interno)
- Prof. Dr. Rafael Ferreira de Mello (rafaelflmello@gmail.com) - DEInfo/UFRPE (avaliador externo)

Assinaturas

Recife, 17 de Abril de 2017

Ihago Henrique Lucena e Silva
(Aluno)

Prof. Dr. Rafael Dueire Lins
(Orientador)