Sumarização Automática para Simplificação Textual

Paulo Rodrigues Alves Margarido, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos
pauloram@grad.icmc.usp.br, taspardo@icmc.usp.br

1. Introdução

Recentemente, no Brasil, tem-se dado muita atenção à questão da inclusão digital. Porém, a maioria das iniciativas criadas visam apenas a fornecer o acesso ao conteúdo digital, mas não procuram adaptar este conteúdo. Pessoas com baixo nível de letramento nem sempre têm condições de compreender o conteúdo. O projeto PorSimples visa criar um sistema de simplificação automática para adaptar conteúdo *online* para pessoas analfabetas ou incapacitadas. Em particular, empregam-se métodos de sumarização para simplificar textos. Sumarização é a tarefa de Processamento de Língua Natural (PLN) que consiste em gerar resumos (sumários) a partir de textos maiores. Os métodos aqui descritos são *extrativos,* isto é, eles montam sumários a partir da extração e justaposição de sentenças completas do texto original.

2. Objetivos

O principal objetivo deste trabalho foi determinar um método de sumarização automática ideal para compor um módulo de sumarização para o sistema final. Vários algoritmos, alguns clássicos e alguns do estado da arte foram testados em busca de um método mais satisfatório. Uma vez que o método de sumarização foi escolhido, foi feito um experimento para averiguar o impacto causado pela sumarização no problema da simplificação com usuários reais.

3. Métodos

Vários métodos de sumarização foram testados. Destes, quatro são baseados em palavras-chave (as sentenças que contêm alguma palavra-chave fazem parte do sumário): EPC-P, EPC-R [3], e suas variações que levam em conta o número de palavras-chave encontradas por sentença; um é o GistSumm, baseado na sentença principal do sumário [1]; três são baseados em aprendizado de máquina: SuPor-2, TextRank e TextRank+Thesaurus [2]; dois são baselines: *First Sentences* escolhe sempre as primeiras sentenças e *Random* as escolhe aleatoriamente.

Os métodos de sumarização foram comparados com base em dois parâmetros: sua capacidade de detectar a sentença principal de um texto e a qualidade dos sumários gerados. Para o primeiro parâmetro, foi utilizado um conjunto de 187 textos jornalísticos, e todos os métodos geraram sumários de uma sentença. Dado o gênero, assumiu-se que a primeira sentença era a mais importante e, para verificar isso, foi feito um teste manual sobre uma amostra aleatória de textos. Para o segundo parâmetro foi utilizada a métrica ROUGE [4] que, dado um sumário automático e um sumário manual (considerado ideal), retorna uma pontuação entre 0 e 1 para o sumário automático.

4. Resultados

Uma vez feita a avaliação descrita na Seção 3, foi construída a Tabela 1 contendo os resultados.

Tabela 1: Performance dos métodos de sumarização

Método	Precisão	Precisão Manual	ROUGE
EPC-P	89%	85%	0.4722
EPC-P2	58%	60%	0.4640
EPC-R	60%	60%	0.4736
EPC-R2	39%	35%	0.4655
GistSumm	46%	50%	0.4185
SuPor-2	75%	85%	0,5839
TextRank	39%	50%	0,5426
TextRank+Thesaurus	27%	35%	0,5603
First Sentences	96%	90%	0.4369
Random	30%	25%	0.3121

Pode-se ver que a performance do método EPC-P é boa em ambos os parâmetros. Os métodos SuPor-2 e TextRank são bons, porém são bastante pesados e pouco portáveis, fazendo do EPC-P a melhor escolha para um sistema com mais módulos.

Com base nestes resultados, foi feita uma avaliação com usuários reais para averiguar o impacto da sumarização no entendimento. 19 pessoas participaram do experimento, e cada uma delas leu um par de textos: o texto original e uma das seguintes: apenas o sumário, o texto original com o sumário em negrito ou o texto original com a sentença principal em negrito.

Das 3 pessoas com até 2 anos de estudo, 2 não terminaram o experimento, e 1 pessoa que terminou o experimento considerou que os destaques não ajudaram.

Das 5 pessoas com até 5 anos de estudo, 66% consideraram que o sumário foi mais fácil de entender; 100% acharam o texto com as sentenças principais em destaque igualmente inteligíveis; e 60% acharam o texto com a sentença principal em destaque mais difícil.

Das 6 pessoas com até 8 anos de estudo: 100% consideraram ler o sumário menos cansativo, mas ambos o texto original e o sumário igualmente inteligíveis; 75% consideraram o texto com as sentenças importantes em destaque mais fácil; e 60% acharam o texto com a sentença principal em destaque mais difícil de entender.

Das 5 pessoas com mais de 10 anos de estudo, 57% acharam o texto mais fácil de entender do que o sumário, mas o sumário menos cansativo; 100% acharam o texto original mais fácil do que o texto com as sentenças importantes em destaque; e 80% consideraram texto com a sentença principal em destaque mais fácil de entender.

5. Conclusões

A partir dos experimentos aqui descritos, concluiu-se que, embora uma forma de simplificação aqui proposta não seja universal, a sumarização colabora de alguma forma na compreensão do texto. Experimentos maiores talvez forneçam resultados mais precisos, mas este experimento piloto demonstra que o sistema tem potencial.

6. Referências Bibliográficas

[1] Mihalcea, R., Tarau, P. TextRank: Bringing Order into Texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2004)

[2] Pardo, T.A.S., Rino, L.H.M., Nunes, M.G.V. GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken - PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218 (2003)

[3] Pereira, M.B., Souza, C.F.R., Nunes, M.G.V. Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português. Revista Eletrônica de Iniciação Científica. Ano II, Volume II, Número I (2002)

[4] Lin, C., Hovy, E.H. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of Language Technology Conference (2003)