Avaliação Comparativa de Métodos de Sumarização Automática Baseados na Teoria de Estruturação Discursiva

Vinícius R. de Uzêda, , Thiago A. S. Pardo, Maria das Graças V. Nunes
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

vruzeda@grad.icmc.usp.br, {taspardo, gracan}@icmc.usp.br

1. Introdução

A disponibilidade de informações vem crescendo rapidamente hoje em dia, e é impossível digerir tamanha quantidade de dados. Para facilitar esse trabalho, as pessoas buscam versões mais compactas, resumidas: sumários. Surgiu então a área de Sumarização Automática (SA), propondo métodos de usar o computador para criar esses resumos.

Existem duas abordagens para tanto: a superficial, que utiliza estatísticas simples; e a profunda, que faz uso de conhecimentos lingüísticos para melhorar a qualidade dos sumários gerados. Acredita-se que existe uma relação entre o nível de complexidade dessas informações lingüísticas, como a sintaxe, a semântica ou o discurso, e a qualidade dos sumários obtidos [1]. Uma das principais teorias que representam o nível de conhecimento discursivo é a Teoria de Estruturação Retórica (do inglês, *Rhetorical Structure Theory* - RST) [2].

2. Objetivos

Vários pesquisadores propuseram diferentes métodos de sumarização automática fazendo uso desta teoria, mas ainda não havia se comparado tais algoritmos. O propósito deste trabalho foi justamente realizar tal avaliação.

3. Métodos

Foram implementados cinco dos métodos mais conhecidos, identificados pelos nomes de seus idealizadores: Marcu [3], e uma versão aperfeiçoada por ele mesmo (aqui denominada MAperf) [4], O'Donnell [5] e Ono [6], além de um novo método proposto neste trabalho (tratado por Uzêda). A idéia por trás de todos eles é atribuir valores aos segmentos textuais expressos na análise/estrutura retórica de um texto segundo a RST. Tais valores represem a importância dos segmentos no texto, sendo que os mais importantes serão selecionados para compor o sumário.

O método proposto é baseado nos demais: a pontuação inicial de cada sentença é igual ao dobro da profundidade da estrutura retórica. Realiza-se, então, um percurso pela estrutura, buscando cada sentença, e, à sua pontuação, desconta-se uma unidade para cada ligação de um segmento menos importante percorrida, e subtrai-se ainda, para cada nível em que a sentença não se encontra no conjunto de elementos mais importantes, um fator de desinteresse de cada relação em que isso ocorre.

Além disso, utilizou-se como método *baseline* na avaliação um sistema superficial, o GistSumm [7], em duas variações, com ou sem a aplicação do método *Average keywords* (tratado como AvgKw).

4. Resultados

Para a avaliação dos métodos, utilizou-se o *RST Discourse Tree Bank* [8], um conjunto de textos manualmente anotados segundo a

RST, já acompanhados de seus sumários humanos.

Foram utilizadas as medidas do pacote ROUGE [9], forma de avaliação automática de sumarizadores. A medida compara os sumários produzidos automaticamente com os sumários humanos correspondentes, atribuindo-lhes notas de 0 a 1, sendo 1 o melhor possível.

A Tabela 1 mostra as medidas obtidas para todos os métodos avaliados. Os métodos baseados na RST estão longe de excelentes, mas superam os superficiais. O método proposto está entre os melhores observados.

Tabela 1. Resultados Obtidos

GistSumm	0,34078	O'Donnell	0,34794
AvgKw	0,33289	Ono	0,39118
Marcu	0,38413	Uzêda	0,38845
Maperf	0,38475		

5. Conclusão

Os métodos de SA baseados em RST apresentam resultados promissores, mas muito ainda precisa ser estudado para que resultados melhores sejam obtidos.

6. Agradecimentos

Às agências de fomento à pesquisa FAPESP, CAPES e CNPq.

7. Referências Bibliográficas

[1] Pardo, T.A.S. (2005). *Métodos para Análise Discursiva Automática*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Junho, 211p.

[2] Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

[3] Marcu, D. (1997). From discourse structures to text summaries. In I. Mani e M. Maybury (Eds.), *Proceedings of ACL Workshop on Intelligent Scalable Text Summarization*, pp. 82-88. Madrid, Spain.

[4] Marcu, D. (1998). To build text summaries of high quality, nuclearity is not sufficient. In the *Working Notes of the the AAAI-98 Spring Symposium on Intelligent Text Summarization*. Stanford, CA.

[5] O'Donnell, M. (1997). Variable-Length On-Line Document Generation. In the *Proceedings of the 6th European Workshop on Natural Language Generation*. Duisburg, Germany.

[6] Ono, K.; Sumita, K.; Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In the *Proceedings of the International Conference on Computational Linguistics (Coling-94)*.

[7] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In the Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal.

[8] Carlson, L.; Marcu, D.; Okurowski, M.E. (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.), *Current Directions in Discourse and Dialogue*, pp. 85-112. Kluwer Academic Publishers.

[9] Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-Gram Co-occurrence Statistics. In the *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003). Edmonton, Canada.