Investigação de Métodos de Segmentação e Agrupamento de Subtópicos para Sumarização Multidocumento

Rafael Ribaldo, Paula C. F. Cardoso, Thiago A. S. Pardo

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brasil

ribaldo@usp.br, paulastm@gmail.com, taspardo@icmc.usp.br

1. Introdução

Nas últimas décadas, observa-se um constante aumento no volume de informação na web. Aplicações computacionais, como a sumarização automática multidocumento, têm se tornado cada vez mais necessárias. Dentre os desafios dessas aplicações encontra-se o tratamento da variedade de subtópicos de uma coleção de textos relacionados. Por exemplo, um texto sobre uma partida de futebol pode descrever a partida, o placar/resultado final e partidas futuras, subtópicos que deveriam estar minimamente representados em um sumário desses textos.

Este projeto visa adaptar e explorar uma técnica clássica para a sumarização multidocumento abordando a questão de subtópicos. Para isso, propõe-se, em primeiro lugar, identificar os subtópicos presentes nos diversos textos de um mesmo grande tópico. O segundo passo visa agrupar os subtópicos de diferentes textos (cenário multidocumento), pois eventualmente pode haver subtópicos semelhantes, uma vez que textos que versam sobre um mesmo assunto (tópico) podem conter passagens muito parecidas. Por fim, usa-se um dos métodos de Salton et al. (1997) para a construção de sumários a partir dos segmentos textuais relevantes detectados.

Na Seção 2 são apresentados os conceitos básicos da área, sendo seguidos pelas atividades a serem desenvolvidas para a conclusão e obtenção de resultados deste projeto, na Seção 3.

2. Conceitos Básicos

2.1. Sumarização

Salton et al. (1997) desenvolveram vários métodos que servem de base para este trabalho. Por esses métodos, um texto é representado em um grafo não direcionado (correspondente ao que Salton et al. chamam de mapa de relacionamentos), sobre o qual os algoritmos de seleção de conteúdo podem ser aplicados. São eles: Caminho Denso (no original, *bushy path*), Caminho Profundo (*depth-first path*) e Caminho Denso Segmentado (*segmented bushy path*). Ainda que os dois primeiros algoritmos resolvam (1) a possível má legibilidade do sumário gerado e (2) a possível falta de coerência, eles causam outro problema que corresponde aos caminhos não cobrirem todos os subtópicos provenientes do texto de origem, sendo necessária a utilização do caminho denso segmentado. O Caminho Denso Segmentado ataca o problema mencionado anteriormente de forma que são construídos diversos caminhos densos para cada subtópico e, logo depois, concatena-os em ordem textual. Garante-se não somente que pelo menos um parágrafo de cada subtópico seja selecionado para compor o sumário,

Anais do III Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana, páginas 25–27, Fortaleza, CE, Brasil, Outubro 21–23, 2013.

mas também algum parágrafo de transição entre subtópicos também seja incluído (respeitando a taxa de compressão), de forma que o sumário detenha de uma melhor legibilidade. Como se pode notar, faz-se necessário a esse caminho de Salton et al. a segmentação dos textos fonte em subtópicos.

2.2. Segmentação Topical

Vários sistemas de sumarização fazem uso de segmentadores topicais, tais como o *TextTiling* (Hearst, 1997). Nesse sistema, assume-se que um conjunto de itens lexicais é usado durante o desenvolvimento de um subtópico e, quando o subtópico muda, uma proporção significante de vocabulário também muda. Para identificar as grandes mudanças de subtópicos, blocos de textos são comparados entre si. Quanto mais palavras em comum os blocos tiverem, maior a chance de fazerem parte do mesmo subtópico.

Por fim, como o segmentador *TextTiling* pertence ao cenário monodocumento, um passo adicional de agrupamento/correlação das passagens de subtópicos semelhantes deve ser considerado, uma vez que este projeto está inserido no cenário multidocumento.

2.3. Agrupamento

Agrupamento (*clustering*) é uma noção que surge naturalmente em muitos campos, sempre que há conjuntos heterogêneos de objetos. É natural a busca por métodos para agrupar tais objetos com base em alguma medida de semelhança. Por exemplo, para definir a distância entre objetos, pode-se considerar que quanto mais perto eles estão entre si, mais parecidos eles são.

Há diversos trabalhos na área, como a ferramenta para a língua portuguesa SiSPI (Seno, 2008), que agrupa sentenças similares utilizando o algoritmo *Single-pass* (Van Rijsbergen, 1979). Tal como o nome sugere, *Single-pass*, no contexto do descobrimento de subtópicos correlacionados, requer uma única passagem sequencial ao longo do conjunto de subtópicos a serem agrupados. É um algoritmo de agrupamento incremental (grupos são criados de forma incremental, analisando-se todos os outros previamente criados). Inicialmente, o algoritmo cria o primeiro grupo, selecionando o primeiro subtópico de uma coleção de documentos a ser agrupado. Então, este primeiro grupo inicia o trabalho de agrupamento com todos os subtópicos remanescentes. Em cada análise dos grupos, o algoritmo decide se um subtópico recém-selecionado deve ser colocado em um grupo já criado ou em um novo. Esta decisão é feita de acordo com uma condição especificada pela função de similaridade empregada, isto é, um limite de similaridade previamente determinado. Neste trabalho, a função de similaridade é a medida de similaridade lexical do cosseno – quanto maior o valor de similaridade entre dois subtópicos, mais semelhantes eles são.

3. Atividades Futuras

Como atividades futuras, espera-se combinar as técnicas de identificação (*TextTiling*) e agrupamento de subtópicos com o caminho denso segmentado (Salton et al., 1997) para selecionar conteúdo relevante para o sumário. Feito isso, serão realizadas análises dos resultados obtidos utilizando o córpus de referência CSTNews (Cardoso et al., 2011) e o pacote de métricas para avalição de sumários ROUGE (*Recall-Oriented Understudy for*

Gisting Evaluation) (Lin e Hovy, 2003), amplamente utilizado na área de sumarização. Os resultados serão comparados com outros trabalhos da área para a língua portuguesa.

Acredita-se que as técnicas utilizadas produzirão sumários de melhor qualidade em relação ao que existe atualmente para a língua portuguesa, pois, com base nelas, realiza-se um processamento textual mais informado.

Agradecimentos

À FAPESP, à CAPES e ao CNPq pelo apoio a este trabalho.

Referências

Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese.In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105.

Hearst, M. (1997).TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages.*Computational Linguistics*, Vol. 23, N. 1, pp. 33-64.

Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In the *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 71-78.

Salton, G.; Singhal A.; Mitra, M; Buckley C. (1997). Automatic Text Structuring And Summarization. *Information Processing & Management*, Vol. 33, No, 2, pp. 193-207.

Seno, E.R.M. and Nunes, M.G.V. (2008).Some Experiments on Clustering Similar Sentences of Texts in Brazilian Portuguese.In the*Proceedings of the International Conference on Computational Processing of Portuguese Language* – PROPOR (Lecture Notes in Artificial Intelligence, 5190), pp. 133-144.

Van Rijsbergen, C.J. (1979). *InformationRetrieval*. Butterworths, Massachusetts.