Estudo de Métodos Clássicos de Sumarização no Cenário Multidocumento Multilíngue

Fabrício Elder da Silva Tosta[1,2], Ariani Di Felippo[1,2], Thiago A. S. Pardo[2]

[1] Departamento de Letras (DL) – Centro de Educação e Ciências Humanas (CECH)
Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905– São Carlos – SP – Brazil

[2] Núcleo Interinstitucional de Linguística Computacional (NILC)
Inst. de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 - São Carlos, - SP – Brazil

fabricio3341@hotmail.com, arianidf@gmail.com, taspardo@icmc.usp.br

Abstract: *In this paper, we present the undergraduate research project that aims at investigating the application of classical superficial methods of automatic summarization in the multilingual multi-document scenario. This project is motivated by the necessity of studies on this task involving Brazilian Portuguese (BP). Consequently, we intend to provide subsidies for the developing of multilingual multi-document summarization systems for BP.*

Resumo: *Neste artigo, apresenta-se o projeto de iniciação científica em que se investiga a aplicação de métodos superficiais clássicos de sumarização automática na tarefa de sumarização multidocumento multilíngue (SAMM). Tal projeto é motivado pela necessidade de estudos nesse cenário que envolvendo o português do Brasil (PB). Com isso, espera-se gerar subsídios para o desenvolvimento de sistemas de SAMM aplicados ao PB.*

1. Introdução

Levando-se em conta a crescente quantidade de informação disponível em diferentes línguas, sobretudo na *web*, e o pouco tempo que as pessoas têm para assimilá-las, a tarefa realizada na subárea do Processamento Automático das Línguas Naturais (PLN) denominada *Sumarização Automática Multidocumento Multilíngue* (SAMM) faz-se útil na medida em que agiliza o acesso à informação. Nessa tarefa, processam-se coleções de textos-fonte que versam sobre um mesmo assunto em duas ou mais línguas e, a partir delas, produzem-se sumários em uma das línguas dos textos-fonte [Mckeown e Radev 1995; Mani 2001]. Quanto à SAMM, não se tem conhecimento sobre pesquisas que envolvem o português do Brasil (PB). Diante desse cenário, objetiva-se investigar alguns métodos clássicos superficiais de sumarização automática monodocumento (SA) no cenário da SAMM envolvendo o PB. Tais métodos, ditos superficiais, utilizam-se pouco ou nenhum conhecimento linguístico para produzir sumários; o conhecimento comumente utilizado é o empírico/estatístico, sendo que os sumarizadores dessa abordagem costumam produzir *extratos*, ou seja, sumários compostos por trechos inalterados do(s) texto(s)-fonte [Sparck e Jones 1993]. Neste texto, em especial, apresentam-se especificamente os métodos a serem investigados e as etapas de realização da pesquisa. Na próxima Seção, descrevem-se os principais métodos

superficiais. Na Seção 3, apresenta-se a metodologia de pesquisa. Por fim, na Seção 4, algumas considerações finais sobre o trabalho são feitas.

2. Breve Revisão Bibliográfica

Na SA, vários métodos superficiais de seleção de conteúdo podem ser aplicados a um texto com o objetivo de: (i) reconhecer as unidades de significado do texto-fonte que contêm as ideias centrais do mesmo, (ii) ranquear tais sentenças e (iii) extrair as mais pontuadas para compor o sumário [Mani 2001]. Dentre os que se pretende investigar, estão: (a) *método das palavras-chave,* que consiste na identificação de palavras-chave, isto é, das mais freqüentes, selecionam-se para compor o sumário as sentenças em que tais palavras ocorrem, até que a taxa de compreensão seja atingida [Black e Johnson 1988]; (b) *método de localização,* que consiste basicamente em classificar as primeiras sentenças do texto-fonte como as mais importantes [Baxendale 1958]; (c) *método de expressões indicativas,* que consiste na seleção das sentenças com base na frequência de certas expressões sinalizadoras (em inglês, *cue words*) [Paice 1981], como "o objetivo deste trabalho é..." "os resultados são", etc.; e (d) *método relacional*, que consiste em inter-relacionar as sentenças do texto-fonte com base nas suas palavras constitutivas e, a partir desse relacionamento, selecioná-las para compor o sumário [Skorokhod´ko 1972].

A seguir, apresentam-se as etapas previstas para a investigação em questão.

3. Metodologia

Além da revisão da literatura sobre a SAMM e os métodos superficiais clássicos de seleção de conteúdo, as seguintes etapas são propostas para a investigação:

(a) Seleção e/ou composição de *corpus*: consiste na seleção e/ou construção de um *corpus* jornalístico multidocumento e multilíngue. Tal *corpus* deve ser composto por uma coleção de textos jornalísticos que versam sobre um mesmo assunto em PB, inglês e espanhol. Caso a construção seja necessária, a *web* será utilizada como fonte e a coleta será feita por meio de algum motor de busca como o Google[1].

(b) Tradução dos textos-fonte: consiste na tradução dos textos-fonte originalmente em inglês e espanhol para PB, pois, assim, o processo de SA será feito a partir de textos em PB. Para tanto, investigar-se-á a utilização de ferramentas automáticas ou serviços on-line gratuitos de tradução como o Google Translator[2].

(c) Segmentação e uniformização dos textos: a segmentação consiste, no caso, em segmentar os textos do *corpus* em sentenças. A uniformização dos dados, que visa melhores resultados de processamento, pode englobar vários processos: (i) *case folding*, em que todas as letras das sentenças são transformadas em minúsculas; (ii) *lematização* (*lemmatizer*) ou *radicalização* (*stemming*), em que as palavras são reduzidas às suas formas canônicas ou a seus radicais, respectivamente; (iii) *remoção de stopwords*, em as palavras funcionais (p.ex.: preposições, artigos, etc) são excluídas.

(d) Pontuação e ranqueamento das sentenças: consiste em atribuir uma pontuação às sentenças identificadas em (c) e produzir um ranque das mesmas. A pontuação será em função dos métodos investigados e, com base na pontuação, as sentenças são ordenadas em um ranque.

---

[1] http://www.google.com.br

[2] http://translate.google.com.br/#

(e) Seleção das unidades textuais e produção dos sumários: consiste na seleção ou eleição das sentenças dos textos-fonte que constituirão o sumário em função de uma taxa de compressão. Para a seleção das sentenças, algumas questões podem ser consideradas, como: (i) combinação de dois ou mais métodos, (ii) especificação de um *threshold* (isto é, um valor mínimo de relevância aceitável, por exemplo, a média das pontuações de todas as sentenças dos textos-fonte), (iii) coerência/coesão entre as sentenças selecionadas e (iv) língua nativa do leitor. Quanto a (iv), caso duas sentenças dos textos-fonte obtenham a mesma pontuação ou pontuações próximas, dar-se-á preferência à sentença proveniente dos textos originais em PB e não das traduções.

(f) Avaliação dos métodos investigados: consiste no julgamento humano dos sumários produzidos em (e) quanto a: (i) informatividade, (ii) a coerência e a coesão, (iii) inadequações linguísticas, (iv) estranhezas provenientes do processo de tradução, (v) falta de adaptações culturais, etc. Tal avaliação é dita intrínseca, pois foca a qualidade do resultado obtido [Sparck Jones e Galliers 1996]. Eventualmente, a avaliação pode indicar novas estratégias e critérios para seleção de informações no cenário multidocumento multilíngue.

4. Considerações Finais

Atualmente, a pesquisa encontra-se na fase de levantamento bibliográfico e seleção dos textos para a composição do *corpus*. Ao final, espera-se gerar subsídios para o desenvolvimento de sistemas de SAMM aplicados ao PB.

5. Referências Bibliográficas

Mani, I. (2001) Automatic Summarization. John Benjamins Publishing Co., Amsterdam.

Mckeown, K. e Radev, D. R (1995) "Generating summaries of multiple news articles". In the Proceedings of the 18th ACM-SIGIR. Seattle, p. 74-82.

Spark Jones, K. (1993) Discourse modeling for Automatic Summarization. Technical. Report, n. 290. University of Cambridge, UK.

Black, W.J. e Johnson, F.C. (1998) Expert systems for information management: a practical evaluation of two rule-based automatic abstraction techniques. University of Manchester: Institute of Science and technology, vol.1, n.3.

Baxendale, P. B. (1958) Machine-made index for technical literature-an experiment. In *IBM Journal of Research and Development*, no. 2, p.354-361.

Paice, C. D. (1981) "The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases". In the Proceedings of the 3rd ACM-SIGIR. Kent, p.171-191.

Skorokhod´ko, E. F. (1972) "Adaptive Method of automatic abstracting and indexing". In the Proceedings of the IFIP. Ljubjana, Yogoslavia, p.449-460.

Sparck Jone, K. (2007) Automatic summarising: a review and discussion of the state of art. Technical Report UCAM-CL-TR-679. University of Cambridge, Ca.

____ e Galliers, J. (1996) Evaluating Natural Language Processing systems: an analysis and review. Springer-Verlag, Berlin.