Estudo de Métodos Clássicos de Sumarização no Cenário Multidocumento Multilíngue

Fabricio E. da S. 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

1. Introdução

Diante da enorme quantidade de informação disponível em diferentes línguas na *web* e do pouco tempo que se têm para assimilá-la, tem-se intensificado o interesse pela *Sumarização Automática Multidocumento Multilíngue* (SAMM). Nela, gera-se, a partir de uma coleção de dois ou mais textos sobre um mesmo tópico em línguas distintas, um único sumário, expresso em uma das línguas dos textos-fonte [Mani 2001]. Idealmente, o sumário deve expressar a informação principal da coleção de forma coesa e coerente.

Em Evans *et al*. (2004), identifica-se um método simples (*baseline*) que estabelece as seguintes etapas para a SAMM: (i) tradução automática (TA) dos textos-fonte da coleção para a língua em que se deseja produzir o sumário; (ii) seleção das sentenças dos textos-fonte (originais e traduções) que expressam a ideia central da coleção com base em critérios linguísticos superficiais, e (iii) produção do sumário pela justaposição das sentenças na ordem em que foram selecionadas. Assim, a SAMM engloba problemas de gramaticalidade gerados pela TA e fenômenos típicos da multiplicidade de textos-fonte, como a redundância de informação.

Iniciando as pesquisas sobre SAMM no Brasil, investigou-se a aplicação manual do método *baseline* para a produção de sumários multidocumento multilíngue em português. Especificamente, foram testados 4 métodos *baseline*, variando-se o critério de seleção de conteúdo e o tratamento da redundância/tradução: Método 1 – Frequência; Método 2 – Localização; Método 3 – Frequência com tratamento da redundância e da tradução, e Método 4 – Localização com tratamento da redundância e da tradução.

Para testar os 4 métodos, construiu-se o CM3News, *corpus* multidocumento trilíngue de textos jornalísticos. O CM3News é composto por 10 coleções de textos. Cada coleção possui 3 notícias sobre mesmo assunto em línguas distintas, isto é, português, inglês e espanhol, os quais foram compilados de forma manual, respectivamente, das versões *online* dos jornais: (i) *A Folha de São Paulo*, (ii) *BBC News*; e (iii) *El país*. Os textos em inglês e em espanhol foram traduzidos para o português através do serviço *online Google Translator*.

Na Seção 2, descreve-se a aplicação dos 4 métodos ao CM3News, cujas coleções, após a TA, passaram a conter 1 texto original em português e 2 textos traduzidos. Na Seção 3, apresenta-se o experimento de avaliação e seus resultados. E, na Seção 4, apresentam-se algumas considerações finais.

Anais do III Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana, páginas 34–36, Fortaleza, CE, Brasil, Outubro 21–23, 2013.

2. Aplicação dos Métodos ao *corpus* CM3News

2.1. Método 1 - Frequência

Dada uma coleção, as sentenças dos textos-fonte receberam uma pontuação resultante da soma da frequência de ocorrência na coleção de suas palavras de classe aberta, a partir da qual foram ranqueadas em ordem decrescente. Assim, o topo do ranque foi ocupado pelas sentenças compostas pelas palavras mais frequentes. A pontuação e o ranqueamento foram feitos por uma funcionalidade do sumarizador *GistSumm* [Pardo, 2005]. A partir do ranque, a seleção manual de conteúdo consistiu em: (i) selecionou-se a 1ª sentença do ranque para iniciar o sumário; (ii) selecionou-se a próxima sentença do ranque; (iii) repetiram-se os passos para as demais sentenças do ranque até que a taxa de compressão desejada de 70% fosse atingida, ou seja, até que as sentenças atingissem 30% do tamanho (em número de palavras) do maior texto da coleção. Os sumários foram produzidos pela justaposição das sentenças na ordem em que foram selecionadas.

2.2. Método 2 - Localização

Com base no critério da localização, as sentenças foram caracterizadas em função da sua posição no texto-fonte da coleção. A primeira sentença de cada um dos 3 textos foi especificada com o atributo "inicio", a última, com o atributo "fim", e as demais, com o atributo "meio", nesta última as sentenças foram elencadas por ondem de ocorrência nos textos. Assim, o topo do ranque foi ocupado pelas sentenças "início", seguidas pelas sentenças "meio" e, por fim, pelas sentenças "fim". Uma vez ranqueadas, as sentenças foram então selecionadas seguindo os mesmos passos utilizados no Método 1.

2.3. Método 3 – Frequência com Tratamento da Redundância e da Tradução

A partir do ranque estabelecido para o Método 1, a seleção manual de conteúdo no Método 3 consistiu em: (i) selecionar a sentença de maior pontuação do ranque para iniciar o sumário; (iii) selecionar a próxima sentença do ranque; (iv) calcular a redundância entre a nova sentença candidata e a sentença já selecionada para o sumário; (v) selecionar a sentença candidata para compor o sumário se esta contiver pouca similaridade com a sentença inicialmente selecionada e não apresentar problemas de TA, (vi) substituir a sentença selecionada não-redundante com problemas de tradução por uma similar proveniente do texto-fonte original em português, (vii) repetir os passos para as demais sentenças do ranque até que a taxa de compressão de 70% fosse atingida. A similaridade, tanto para eliminar a redundância como para substituir sentenças traduzidas agramaticais por originais em português, foi calculada de forma automática com base na medida estatística *word overlap*, que se baseia na sobreposição das palavras de classe aberta idênticas [Jurafsky, Martin, 2001]. A produção dos sumários foi manual pela justaposição das sentenças na ordem em que foram selecionadas.

2.4. Método 4 – Localização com Tratamento da Redundância e da Tradução

Com base no ranque produzido para o Método 2, a seleção manual de conteúdo no Método 4 seguiu os mesmos passos do Método 3, já que englobou o tratamento da redundância e dos problemas gerados pela TA. A produção dos sumários também foi feita manualmente pela justaposição das sentenças na ordem em que foram selecionadas.

3. Experimento de Avaliação e seus Resultados

Para verificar o desempenho dos 4 métodos, optou-se por uma avaliação intrínseca (humana), que consistiu na análise da legibilidade (ou fluência) dos sumários gerados para 5 das 10 coleções do CM3News. Os sumários foram analisados por 1 especialista em função dos 5 parâmetros utilizados na DUC (*Document Understanding Conference*)[1]: (i) gramaticalidade, (ii) não-redundância, (iii) clareza referencial, (iv) foco temático, e (v) coesão/coerência. Esses parâmetros foram pontuados com valores de 5 a 1, sendo 5=muito bom, 4=bom, 3=aceitável, 2=ruim e 1=muito ruim. Na Tabela 1, esquematiza-se a média obtida por cada método.

Tabela 1. Média das pontuações dos métodos.

Critérios	Método 1	Método 2	Método 3	Método 4
Gramaticalidade	2	2,3	**3**	2,8
Não-redundância	2	2,8	**3**	3
Clareza referencial	2,8	3	**3,2**	3
Foco temático	4	3,8	**4**	3,8
Coesão e coerência	2,8	2,8	**2,8**	2,4

Pela Tabela 1, vê-se que o Método 3, pautado na localização com tratamento da redundância e da tradução, obteve em média as mais altas pontuações quanto aos 5 parâmetros. Esse resultado não é surpreendente, pois é notória a pertinência da localização na seleção de conteúdo para a produção de sumários jornalísticos.

4. Considerações Finais

Neste trabalho, construiu-se o primeiro *corpus* multidocumento multilíngue que abrange o português e testaram-se 4 métodos *baseline* de SAMM. Atualmente, o *corpus* CM3News está em ampliação e métodos baseados em conhecimento linguístico profundo estão sendo investigados. Reconhece-se que uma limitação deste trabalho reside no fato de a avaliação ter sido feita com base em poucas coleções e por apenas 1 especialista. Assim, no futuro, objetiva-se refinar o processo de avaliação.

References

Evans, D.K.; Klavans, J.L.; McKeown, K.R. (2004). Columbia NewsBlaster: Multilingual news summarization on the web. In the *Proceedings of the HLT/NAACL: Demonstration Papers at HLT-NAACL*, p.1-4.

Gupta, V; Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, Oulu, v. 2, n. 3, p. 258-268.

Jurafsky, D; Martin, J. H. (2007). *Speech and Language Processing*: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 1024P.

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

Pardo, T.A.S. (2005). GistSumm – GIST SUMMarizer: extensões e novas funcionalidades. *Série de Relatórios do NILC*. *NILC-TR-05-05*. São Carlos-SP, 8p.

[1] Em 2008, a DUC tornou-se parte de outra conferência, denominada *Text Analysis Conference* (TAC). O site http://duc.nist.gov/ engloba as informações referentes à DUC de 2001 a 2007.