Enriquecendo o *corpus* CM2News: *Construção e Anotação de Coleções Bilíngues de Notícias*

Yasmin V. Camargo[1,2], Ariani Di-Felippo[1,2]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)

[2]Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

yvizeu@gmail.com, arianidf@gmail.com

Abstract. *We present the extension of CM2News, a multi-document bilingual (Portuguese-English) corpus for Multilingual Multi-document Summarization (MMDS). We added 10 bilingual clusters of news texts to the original set of 20 clusters, and performed a lexical-conceptual annotation of the 20 new source-texts based on WordNet.*

Resumo. *Apresenta-se a extensão do CM2News, corpus multidocumento bilíngue (português-inglês) para pesquisas em Sumarização Automática Multidocumento Multilíngue (SAMM). A extensão consistiu na adição de 10 coleções bilíngues de notícias às 20 pré-existentes e na anotação léxico-conceitual dos 20 novos textos-fonte baseada na WordNet de Princeton.*

1. Introdução

Na Sumarização Automática Multidocumento Multilíngue (SAMM), busca-se, por exemplo, construir métodos que partem de um conjunto de ao menos 2 textos que abordam o mesmo assunto, sendo 1 texto em uma *Lx* e 1 em uma língua *Ly,* e geram um sumário em uma das línguas-fonte (*Lx* ou *Ly*) [Evans *et al* 2004]. Na SAMM, os *corpora* [Berber Sardinha 2004], são recursos centrais ao permitirem modelar computacionalmente a sumarização, além de treinar e avaliar tais modelos/sistemas.

O CM2News [Di-Felippo 2016] é um *corpus* multidocumento bilíngue (português (pt) e inglês (in)) de textos jornalísticos. Sua primeira versão engloba 20 *cluster*, distribuídos em 6 categorias (mundo (8), política (3), saúde (4), ciência (3), entretenimento (1) e meio ambiente (1)). Cada *cluster* contém (i) 2 textos-fonte (1-pt e 1-in), (ii) 1 sumário multidocumento de referência em pt e (iii) anotação manual dos textos-fonte em nível léxico-conceitual. Totalizando 40 textos e 19.983 palavras, o CM2News subsidiou o desenvolvimento de 2 métodos de SAMM envolvendo o português [Tosta 2014, Di-Felippo *et al.* 2016] e o estudo de métricas conceituais (p.ex.: *concept frequency (cf)* e *cf-idf*) que capturam a relevância das sentenças em *clusters* multidocumento multilíngue [Chaud 2015, Chaud e Di-Felippo 2018].

Dada sua relevância/potencial, tem-se focado na expansão do *corpus*, que engloba: (i) construção e anotação léxico-conceitual de novos *clusters* bilíngues, (ii) inclusão de 1 notícia em outra língua estrangeira a todos os *clusters,* e (iii) produção de novos sumários de referência. Na Seção 2, apresenta-se a construção de 10 novos *clusters* bilíngues e, na Seção 3, descreve-se a anotação léxico-conceitual dos novos 20 textos-fonte. Na seção 4, tecem-se algumas observações finais sobre o *corpus*, destacando pesquisas de SAMM em andamento que utilizam sua versão estendida.

2. A construção dos novos *clusters* bilíngue

Os novos *clusters* foram construídos com base nas diretrizes de Tosta [2014] e Di-Felippo [2016], a saber: (i) compilação manual dos textos, (ii) seleção de fontes jornalísticas confiáveis, (iii) compilação de notícias atuais cujos assuntos sejam variados e (iv) seleção de notícias bilíngues de tamanho similar. Assim, as notícias constitutivas dos 10 *clusters* adicionais foram manualmente compiladas das fontes: (i) jornal *A Folha de São Paulo* e portal UOL para os textos em português e (ii) portais *BBC News* e CNN para os textos em inglês. Selecionaram-se notícias publicadas entre abril e outubro de 2018. Com relação à variedade de assuntos, os novos *clusters* distribuem-se nas categorias: saúde (2), poder (1), meio ambiente (3), ciência (1) e entretenimento (3). Sobre a extensão dos textos, ressalta-se que, no geral, os textos-fonte têm tamanho relativamente similares. Os novos *clusters* do CM2News estão descritos na Tabela 1.

Tabela 1. Descrição dos 10 novos *clusters* do CM2News.

Cluster	Domínio	Assunto	Documento	Publicação (data/hora)	Qt. pal./doc	Qt. pal./*cluster*
C21	Poder	Encontro de líderes das Coreias	D1_C21_folha	27/04/18 - 00:34	386	770
			D2_C21_bbc	27/04/18 - 08:06 (GMT)	384	
C22	Ciência	Reprodução de camundongos	D1_C22_folha	11/10/18 - 12:00	578	1.240
			D2_C22_bbc	11/10/18 - 9:46 (GMT)	662	
C23	Entreten.	Kanye West na política	D1_C23_uol	30/10/18 - 19:49	328	782
			D2_C23_bbc	31/10/18 - 7:57	454	
C24	Entreten.	Bebê de Hilary Duff	D1_C24_folha	30/10/18 - 11:00	182	285
			D2_C24_cnn	30/10/18 - 15:21 (GTM)	103	
C25	Entreten.	Acusações a Stallone	D1_C25_uol	31/10/18 - 05:05	150	280
			D2_C25_bbc	31/10/18 - 1:45 (GTM)	130	
C26	Meio ambiente	Oleoduto EUA-Canadá	D1_C26_folha	09/11/18 - 17:55	428	973
			D2_C26_bbc	09/11/18 - 14:32 (GTM)	545	
C27	Meio ambiente	Ataque de leoa em zoológico	D1_C27_folha	22/10/18 - 16:15	220	419
			D2_C27_bbc	22/10/18 - 10:11 (GMT)	199	
C28	Meio ambiente	Baleia morta na Indonésia	D1_C28_uol	21/11/18 - 11:21	287	646
			D2_C28_cnn	21/11/18 - 16:57 (GMT)	359	
C29	Saúde	EUA poliomielite	D1_C29_folha	17/10/18 - 8:00	390	782
			D2_C29_cnn	23/10/18 - 12:27 (GMT)	392	
C30	Saúde	Camisinha autolubrificante	D1_C30_uol	18/10/18 - 12:17	522	1.056
			D2_C30_cnn	19/10/18 - 15:20 (GMT)	434	
Total						7.233

3. A anotação léxico-conceitual dos novos textos-fonte

A anotação foi feita por um 1 linguista computacional e durou 20 dias, em sessões diárias de 60 a 90 minutos. Seguindo Di-Felippo [2016], explicitaram-se os conceitos nominais via MulSen[1] (versão multilíngue do NASP [Nóbrega 2013]), que (i) identifica os nomes via *tagging* (etiquetação morfossintática) e (ii) recupera os conceitos da WordNet de Princeton [Fellbaum 1998] em função do nome a ser anotado. A anotação seguiu 4 regras gerais: (i) anotar de início o texto em inglês do *cluster*, pois a recuperação dos possíveis conceitos da WordNet é direta, (ii) anotar os nomes não

[1] http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/files/MulSEN.zip.

detectados pelo *tagger*, (iii) ignorar as palavras erroneamente detectadas como nome pelo *tagger*, e (iv) anotar as ocorrências de um conceito com o mesmo *synset*[2].

A anotação léxico-conceitual de um nome *n* em inglês inicia com a recuperação automática de todos os *synsets* que contêm *n* e a sugestão do *synset* mais adequado por um algoritmo de desambiguação [Nóbrega 2013], que pode (ou não) ser validado.

Caso a sugestão não seja adequada, pode-se identificar outro *synset* entre os recuperados pelo MulSen. A anotação de um nome *n* em português segue basicamente os mesmos passos. A exceção é um processo adicional de tradução automática (pt-in) (via WordReference®[3]) para que o MulSen recupere os conceitos/*synsets* na WordNet em função do nome que se quer anotar.

A anotação conceitual seguiu 4 regras específicas: (i) uma vez o *tagger* identifica apenas lexias simples (e.ex: [gás_N de pimenta]), anotar todo nome que é núcleo de uma expressão multipalavra com o *synset* que expressa o significado da expressão (p.ex.: [gás<{pepper spray}> de pimenta]; (ii) analisar todas as traduções recuperadas do WordReference® antes de selecionar a mais adequada, posto que a melhor tradução não necessariamente é a primeira da lista recuperada pelo editor; o mesmo procedimento se aplica à seleção do *synset*; (iii) caso necessário, procurar traduções mais adequadas em repositórios externos, inserindo-as manualmente no MulSen, e analisar todos os *synsets* recuperados em função dessas traduções, e (iv) caso a WordNet não contiver certo conceito, selecionar um *synset* hiperônimo, pois os conceitos nominais estão organizados hierarquicamente na base de dados. No total, anotaram-se 1.593 nomes, distribuídos nos *clusters* como indicado na Tabela 2.

Tabela 2. Distribuição da anotação léxico-conceitual nos novos *clusters* bilíngues do CM2News.

Cluster	Qt. Nomes anotados	Cluster	Qt. Nomes anotados
C21	202	C26	163
C22	255	C27	87
C23	143	C28	134
C24	68	C29	250
C25	51	C30	240

4. Considerações finais

Por meio da expansão aqui descrita, o CM2News passou a ter 30 coleções bilíngues (pt-en) que totalizam 27.270 palavras, aumentando o volume de dados para as pesquisas em SAMM que envolvem o português. Atualmente, tem-se conduzido a produção de sumários (abstrativos) de referência em português para os novos *clusters,* seguindo os critérios de Tosta [2014]. Com base na anotação léxico-conceitual dos nomes nas 30 coleções, Camargo [2018] tem conduzido o desenvolvimento de métodos de seleção de conteúdo para a SAMM que consideram (i) pontuação ou peso diferenciado para os conceitos mais genéricos (hiperônimos) de um *cluster*, os quais são relevante para a construção de extratos do tipo informativo/genérico, e (ii) identificação da redundância baseada na medida *concept overlap,* que considera a ocorrência de expressões distintas de um mesmo conceito (sinonímia e equivalência) no *cluster* para calcular a similaridade entre sentenças. Nascimento [2018], por sua vez, conduz uma pesquisa que visa refinar a avaliação de métodos extrativos de SAMM, variando (i) a taxa de

[2] Conjunto de sinônimos que representa um conceito lexicalizado; p.ex: o conceito "veículo que se move por motor próprio" é representado pelo *synset* {car, auto, automobile, machine, motorcar}.
[3] http://www.wordreference.com/

compressão (isto é, tamanho ou extensão em número de palavras) dos extratos automáticos e (ii) a língua materna dos produtores dos sumários de referência. Para tanto, tem-se adicionado 1 notícia em alemão a cada um dos 30 *clusters,* transformando-os em coleções trilíngues, e construído sumários de referência (em português) produzidos por falantes do português e do inglês a partir da leitura dos 3 textos-fonte.

Agradecimento. À CAPES, pelo suporte financeiro.

Referências bibliográficas

Berber Sardinha, T. B. (2004). Lingüística de corpus. São Paulo, Manole, 410 p.

Camargo, Y.V. (2018). Multilingual Multi-Document Summarization: content selection and redundancy treatment based on lexical-conceptual knowledge. In the Proceedings of the Student Research Workshop (SRW) of the 13th International Conference on the Computational Processing of Portuguese, pp. 1-4. September, 24, Canela/RS, Brazil.

Chaud, M.R. (2015). Investigação de estratégias de seleção de conteúdo baseadas na UNL (Universal Networking Language). 2015. 157f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, SP.

Chaud, M.R.; Di-Felippo, A. (2018) Exploring content selection strategies for Multilingual Multi-Document Summarization based on the Universal Network Language (UNL). Revista de Estudos da Linguagem, v. 26 (1), p. 45-71.

Di-Felippo, A. (2016). CM2News: Towards a Corpus for Multilingual Multi-document Summarization. In the Proceedings of the Workshop on Corpora and Tools for Processing Corpora (CTPC), Collocated with PROPOR 2016 (The 12th International Conference on the Computational Processing of Portuguese), Tomar, Portugal, p.1-8.

Di-Felippo, A. Tosta, F. E. S., Pardo, T. A. S. (2016). Applying Lexical-Conceptual Knowledge for Multilingual Multi-Document Summarization. In the Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR). Lecture Notes in Computer Science, Vol 9727, Springer, pp. 38-49, July, 13-15. Tomar, Portugal. ISBN 978-3-319-41552-9

Evans, D.K.; Klavans, J.L.; Mckeown, K.R. (2004). Columbia NewsBlaster: multilingual news summarization on the web. In the Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Boston, p.1-4

Fellbaum, C. (1998): Wordnet: an electronic lexical database (Language, speech and communication). Massachusetts: MIT Press.

Nascimento, D.X. Exploring the evaluation of automatic multilingual multi-document sumaries. In the Proceedings of the Student Research Workshop (SRW) of the 13th International Conference on the Computational Processing of Portuguese, pp. 1-4. September, 24, Canela/RS, Brazil.

Nóbrega, F.A.A. (2013). Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - ICMC, USP, São Carlos.

Tosta, F.E.S. (2014). Aplicação de conhecimento léxico-conceitual na Sumarização Multidocumento Multilíngue. 116p. Dissertação (Mestrado) - Universidade Federal de São Carlos - UFSCar.