ALINHAMENTO MANUAL DE TEXTOS E SUMÁRIOS EM UM *CORPUS* JORNALÍSTICO MULTIDOCUMENTO

1. Introdução

Com o imenso volume de informação disponível na *web*, necessita-se de estratégias que permitam absorvê-la de modo rápido e prático. Assim, a Sumarização Automática Multidocumento (SAM) tem ocupado lugar de centralidade no Processamento Automático de Línguas Naturais (PLN), pois, na SAM, objetiva-se produzir sumários a partir de uma coleção de textos que tratam do mesmo assunto (Mani, 2001).

Nessa subárea, o alinhamento de textos-fonte a seu sumário humano permite, por exemplo, uma análise linguística da Sumarização Humana Multidocumento (SHM), a qual pode gerar regras e modelos explícitos com potencial para subsidiar métodos de SAM mais linguisticamente motivados. O alinhamento (ou anotação) consiste em, dada uma coleção de textos-fonte, relacionar porções de um ou mais textos a seu sumário.

Visando contribuir com a SAM, este artigo apresenta o processo de alinhamento manual de textos-fonte a seus respectivos sumários humanos. Para tanto, utilizou-se o CSTNews (Cardoso *et al*., 2011), que é um *corpus* multidocumento em português do Brasil. Na Seção 2, apresentam-se os pressupostos teóricos que embasam este trabalho. Na Seção 3, citam-se as características do CSTNews. Na Seção 4, elucida-se o processo de alinhamento em questão. Na Seção 5, apresentam-se os resultados do referido alinhamento e, por fim, na Seção 6, algumas considerações finais são feitas.

2. Pressupostos Teóricos

Como mencionado, o alinhamento é um processo que consiste no relacionamento de segmentos textuais com base no conteúdo por eles veiculado. Os segmentos textuais a serem relacionados podem ser palavras, sentenças, parágrafos, seções e até documentos inteiros. O alinhamento é uma tarefa que se iniciou na Tradução Automática, subárea do PLN em que são alinhadas palavras ou sentenças de uma língua-fonte às suas versões em uma língua-alvo (por exemplo, Gale e Church, 1991, 1993; Yamada e Knight, 2001). Um exemplo de alinhamento na tradução pode ser visto no Quadro 1.

Quadro 1: Exemplo de alinhamento na tradução retirado de Gale e Church (1991).

English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Na primeira linha do Quadro 1, 2 sentenças em inglês são alinhadas a 2 sentenças em francês. Na segunda e terceira linhas, 1 sentença em inglês é alinhada a 1 sentença em francês. Já na quarta linha, duas sentenças em inglês são alinhadas a uma única em francês.

Na Sumarização Automática (SA), foco deste trabalho, alinham-se grupos de palavras ou sentenças do(s) texto(s)-fonte(s) a seu sumário (p. ex.: Marcu, 1999; Hirao *et al*., 2004). No Quadro 2, apresenta-se um exemplo de alinhamento na SAM, no qual uma sentença do sumário foi alinhada a duas sentenças, cada uma delas de um texto-fonte distinto.

Quadro 2: Exemplo de alinhamento na SAM.

Sumário	**Documentos**
O Brasil não fará parte do trajeto de 20 países do revezamento da tocha.	A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico.
	O Brasil não faz parte do trajeto da tocha olímpica.

Os primeiros trabalhos sobre alinhamento na SA datam de 1999. Banko *et al*. (1999) e Marcu (1999) propuseram métodos baseados na ocorrência simultânea de palavras idênticas em um texto-fonte e seu sumário (monodocumento), independentemente da ordem de ocorrência das palavras. Os métodos de Jing e McKeown (1999) e Daumé e Marcu (2004, 2005) baseiam-se na utilização de um *Modelo Oculto de Markov* (Baum, 1972). Jing e McKeown (1999) utilizaram heurísticas de operações do tipo *cut and paste* para identificar a origem dos segmentos de um sumário no texto-fonte. Em Daumé e Marcu (2004, 2005), o alinhamento pauta-se em uma *história gerativa* de sumarização humana. Hirao *et al*. (2004), em especial, focam pela primeira vez o alinhamento na SAM. Nesse trabalho, o alinhamento pauta-se na similaridade entre as árvores de dependência dos textos-fonte de uma coleção e a do respectivo sumário multidocumento.

Na Seção 3, apresenta-se o *corpus* utilizado neste trabalho.

3. O *Corpus* CSTNews

O CSTNews (Cardoso *et al*., 2011) é um *corpus* jornalístico com 50 coleções de textos. Cada coleção contém 2 ou 3 textos de diferentes fontes sobre uma notícia e seus respectivos sumários humanos e automáticos, além de diversas anotações. As notícias foram coletadas dos jornais *online Folha de São Paulo, Estadão, Jornal do Brasil, O Globo* e *Gazeta do Povo*, de agosto a setembro de 2007. As coleções foram rotuladas pelas "seções" dos jornais de origem, a saber; "esporte", "mundo", "dinheiro", "política", "ciência" e "cotidiano". O *corpus* possui em média 42 sentenças por coleção (de 10 a 89 sentenças) e em média 7 sentenças por sumário multidocumento (de 3 a 14 sentenças).

4. Análise de *Corpus*

A análise de *corpus* refere-se ao alinhamento manual das sentenças dos textos-fonte das diversas coleções do CSTNews às sentenças dos sumários multidocumento dessas coleções. O alinhamento foi realizado por 2 anotadores da área de Linguística Computacional, cada qual responsável por alinhar metade das coleções.

Antes do alinhamento, realizou-se uma fase de treinamento, na qual 2 coleções foram aleatoriamente selecionadas e alinhadas por cada um dos anotadores,

individualmente, com base na sobreposição de conteúdo (e não na sobreposição lexical). Na sequência, os alinhamentos foram comparados e os casos de divergência foram discutidos com o intuito de ajustar a concordância entre os linguistas computacionais.

Após o treinamento, o alinhamento efetivo dos textos-fonte a seus sumários passou a ser feito em reuniões diárias de 1 ou 2 horas durante o período de aproximadamente 60 dias. Para os casos de alinhamentos duvidosos, porém regulares, criaram-se regras específicas após a análise em conjunto dos mesmos. Tais regras compõem, atualmente, o documento "*Manual de Alinhamento do Corpus CSTNews*".

Para garantir o consenso entre os anotadores, analisou-se a concordância entre eles nas últimas 5 semanas de alinhamento. A cada semana, os pesquisadores alinhavam individualmente uma mesma coleção e comparavam os resultados de cada alinhamento para verificar a concordância. No total, 5 coleções rotuladas por "mundo", "esporte", "dinheiro", "cotidiano" e "política" foram utilizadas nessa tarefa. Da comparação entre os alinhamentos individuais, gerava-se um terceiro alinhamento, resultante do consenso entre os pesquisadores. Os alinhamentos individuais e os consensuais compõem, naturalmente, o conjunto de dados produzidos por este trabalho. Entretanto, apenas os consensuais foram considerados para a análise final dos resultados do alinhamento, os quais serão demonstrados na próxima Seção.

5. Resultados

Como resultado, aproximadamente 70% das sentenças dos sumários multidocumento foram alinhadas a mais de uma sentença dos textos-fonte. Tal fato justifica-se por se tratar de sumários multidocumento, ou seja, versões condensadas de coleções de textos. Todos os tipos de alinhamento resultantes podem ser vistos na Tabela 1.

Tabela 1: Tipos de alinhamento no *corpus*.

Quantidade	**Tipos de alinhamento**												
	1-0	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12
	2	71	90	67	36	37	13	5	5	1	1	2	1

Na Tabela 1, observa-se que: (i) na primeira coluna, 2 sentenças dos sumários não foram alinhadas (1-0), o que resulta da inserção de informação no sumário que não está presente nos textos-fonte; (ii) na segunda coluna, 71 sentenças dos sumários foram alinhadas a 1 sentença dos textos-fonte (1-1); (iii) na terceira coluna, 90 sentenças dos sumários foram alinhadas a 2 sentenças dos textos-fonte (1-2), e assim por diante.

Das 2067 sentenças que compõem os textos-fonte das coleções, 877 (42,43%) foram alinhadas. Porém, isso não significa que estas foram alinhadas uma única vez, já que uma sentença de um sumário pode ser alinhada a mais de uma sentença dos textos-fonte e as sentenças dos textos-fonte podem ser redundantes ou até idênticas.

Quanto à forma de disponibilização, o alinhamento ora apresentado segue o formato de anotação XML (*Extensible Markup Language*), ilustrado no Quadro 3. No Quadro, ilustra-se o alinhamento dos textos-fonte e do sumário multidocumento da coleção 31 do CSTNews. O primeiro bloco do esquema XML descreve o alinhamento da sentença 1 do sumário (de <align SENT="1"> até </align>). Nesse bloco, a sentença 1 (SENT="1") foi alinhada à: (i) SENT="1" do D(ocumento)1, (ii) SENT="1" do D2 e (iii) SENT="2" do D2. Além da informação sobre a sentença e o texto-fonte, o esquema XML prevê a especificação do tipo de alinhamento (no caso, com ou sem contradição) (TYPE="none") e do anotador (JUDGE="veronica").

Quadro 3: Exemplo em XML.

```
<align SENT="1">
        <DOC="D1_C31_Folha.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
</align>
<align SENT="2">
        <DOC="D1_C31_Folha.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="6" TYPE="none" JUDGE="veronica"/>
</align>
<align SENT="3">
        <DOC="D1_C31_Folha.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
        <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
</align>
```

6. Considerações Finais

Destaca-se que o alinhamento dos textos-fonte e sumários multidocumento do CSTNews será utilizado em pelo menos dois trabalhos que se enquadram no cenário da SAM. Em um deles, de natureza computacional, o alinhamento manual será utilizado para avaliar o desempenho de uma ferramenta computacional (isto é, um alinhador automático) que buscará simular o processo humano ora descrito. Tal ferramenta está sendo desenvolvida porque o alinhamento, visto como uma etapa do processo de sumarização, permite a aplicação de diferentes métodos de SAM. Na outra pesquisa, de natureza linguística, o alinhamento manual do CSTNews será o ponto de partida para a investigação de estratégias de SHM que, uma vez formalizadas, possam subsidiar métodos de SAM mais linguisticamente motivados. A análise linguística dos alinhamentos manuais pode revelar estratégias de SHM quando à seleção de conteúdo e produção dos sumários.

Quanto às dificuldades encontradas neste trabalho, destaca-se a necessidade de conhecimento de domínio para realizar o alinhamento dos textos e sumários de certas coleções, sobretudo dos que compõem as coleções rotuladas por “política”.

Referências

Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, v. 3, pp. 1-8.

Banko, M.; Mittal, V.; Kantrowitz, M.; Goldstein, J. (1999). Generating Extraction-Based Summaries from Hand-Written Summaries by Aligning Text Spans. In the *Proceedings of the 4th Conference of the Pacific Association for Computational Linguistics*, 5p.

Cardoso, P.C.F.; Maziero, E.G.; Castro 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.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, v. 22, n. 2, pp. 249-254.

Daumé III, H. and Marcu, D. (2004). A Phrase-Based HMM Approach to Document/Abstract Alignment. In the *Empirical Methods in Natural Language Processing (EMNLP)*, 8p.

Daumé III, H. and Marcu, D. (2005). Induction of Word and Phrase Alignments for Automatic Document Summarization. *Computational Linguistics*, v. 31, n. 4, pp. 505-530.

Gale, W.A. and Church, K.W. (1991). A program for aligning sentences in bilingual corpora. In the *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berkley, pp. 177-184.

Gale, W.A. and Church, K.W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, v. 19, n. 3, pp. 75-102.

Hirao, T.; Suzuki, J.; Isozaki, H.; Maeda, E. (2004). Dependency-based Sentence Alignment for Multiple Document Summarization. In the *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*, pp. 446-452.

Jing, H.; McKeown, K. (1999). The Decomposition of Human-Written Summary Sentences. In the *Proceedings of the 22nd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, pp. 129–136.

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

Marcu, D. (1999). The automatic construction of large-scale corpora for summarization research. In the *Proceedings of the 22nd Conference on Research and Development in Information Retrieval*, pp. 137-144.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In the *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 523-530.