ORDENAÇÃO DE SENTENÇAS EM SUMÁRIOS MULTIDOCUMENTO: UMA ABORDAGEM UTILIZANDO RELAÇÕES DISCURSIVAS

Jader Bruno Pereira Lima, Thiago Alexandre Salgueiro Pardo

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)

jaderlima@grad.icmc.usp.br, taspardo@icmc.usp.br

Abstract. *This paper presents some methods and experiments for sentence ordering in multi-document summaries. In particular, it is proposed a method based on the semantic-discursive theory CST (Cross-document Structure Theory).*

Resumo. *Apresentam-se, neste artigo, alguns métodos e experimentos para a ordenação de sentenças em sumários multidocumento. Em particular, é proposto um método baseado na teoria semântico-discursiva CST (Cross-document Structure Theory).*

1. Introdução

Com a invenção e a popularização da internet, ganhamos uma poderosa fonte de informação e conhecimento, de onde podemos retirar facilmente todas as informações que possam ser relevantes para nossa vida. Essas informações estão dispostas, em sua grande maioria, na forma textual, e, com isso, faz-se cada vez mais necessário o desenvolvimento de ferramentas que auxiliem a manipulação desses dados. Com essa motivação, está sendo desenvolvida no NILC (Núcleo Interinstitucional de Linguística Computacional) uma ferramenta que, a partir de um conjunto de textos jornalísticos que versam sobre determinado assunto, criará um resumo, ou sumário, contendo as informações mais relevantes para o leitor. Tal ferramenta é chamada de sumarizador multidocumento, pois produz um sumário a partir de um conjunto de textos (Mani, 2001). Para o desenvolvimento dessa ferramenta, os trabalhos estão sendo baseados na teoria discursiva multidocumento CST (*Cross-document Structure Theory*) (Radev, 2000), capaz de representar o conteúdo multidocumento, relacionando as várias partes dos textos a serem sumarizados.

Depois de realizada a tarefa de sumarização, as sentenças desses sumários gerados automaticamente necessitam de um processamento final antes de serem apresentadas ao usuário, e uma das principais tarefas é a ordenação das sentenças do sumário, pois a ordem da narração dos fatos/eventos influencia diretamente na coerência e coesão dos sumários, principalmente neste cenário em que as sentenças provêm de diversos textos diferentes. Para justificar a necessidade da ordenação de sentenças, um estudo realizado com juízes humanos (Barzilay et al., 2002) classificou 10 sumários multidocumento gerados automaticamente em três níveis: incompreensíveis, parcialmente compreensíveis ou compreensíveis. Com esse estudo, pôde ser notada a importância da ordem das sentenças no sumário: há uma melhoria considerável na taxa de sumários considerados compreensíveis depois de serem ordenados.

A proposta deste trabalho em andamento é investigar métodos de ordenação, utilizando, como base, as relações CST presentes nos textos. Acredita-se, como hipóteses, que (i) é possível melhorar a compreensão de sumários gerados automaticamente, ordenando as sentenças deste, utilizando a teoria CST como fonte de informação semântica para essa tarefa, e (ii) o uso da CST pode avançar o estado da arte na área. Na seção a seguir são apresentados alguns trabalhos sobre ordenação de sentenças. Na Seção 3 são explicados os métodos desenvolvidos neste trabalho e a etapa atual em que se está.

2. Trabalhos Correlatos

Existem vários trabalhos na linha de ordenação de sentenças de sumários multidocumento. Barzilay et al. (2002) apresentam três métodos de ordenação: a ordenação cronológica, que ordena o sumário considerando apenas a data de criação do texto fonte de cada sentença; a ordenação por maioria, que agrupa as sentenças com o mesmo significado em clusters e então ordena esses clusters de acordo com a posição relativa de cada sentença pertencente ao cluster; e a ordenação aumentada, que faz uma abordagem hibrida combinando os dois métodos anteriores. Okazaki et al. (2004) nos mostram uma outra abordagem para o problema de ordenação de sentenças de sumários, utilizando relações de antecedência entre sentenças. Lapata (2003), utilizando uma abordagem probabilística, apresenta-nos um método que ordena as sentenças levando em consideração a probabilidade condicional de essa sentença ocorrer em determinada posição nos textos de origem.

Aleixo e Pardo (2008) criaram um córpus para fins de sumarização automática multidocumento chamado CSTNews. Esse córpus contém 50 grupos de textos jornalísticos retirados dos principais jornais do país, juntamente com seus sumários feitos manual e automaticamente, os quais servirão de base para testes dos métodos estudados nesse projeto. Os sumários automáticos foram gerados pelo sumarizador *CSTSumm* (Jorge e Pardo, 2010), que também foi desenvolvido no NILC. Para a avaliação dos resultados, foi criada manualmente uma versão alternativa dos sumários automáticos do córpus. Essa versão conta com as sentenças do sumário dispostas com a melhor ordenação possível e seus dados serão usados como referência neste trabalho.

3. Métodos

Três métodos *baseline* foram inicialmente propostos, e, futuramente, terão seus dados comparados com métodos mais complexos que serão apresentados a seguir. Os primeiros métodos ordenam as sentenças do sumário de acordo com o tamanho delas em caracteres: um em ordem crescente, e o outro em ordem decrescente. O terceiro método levou em conta a posição da sentença em seu documento fonte: as sentenças com a posição mais próxima do inicio do texto são colocadas à frente das sentenças mais próximas do final em seu texto fonte, e, em caso de empate, o desempate é feito aleatoriamente.

Para avaliar de maneira exata esses três métodos, foram utilizadas duas medidas: o coeficiente de correlação de Spearman e o coeficiente de correlação de Kendall (Okazaki et al., 2004). Os dois resultados possuem 1 como melhor valor possível, quando as sentenças ordenadas pelo método estão exatamente na posição do sumário ordenado manualmente, e -1 como pior resultado, quando as sentenças estão totalmente invertidas. Os métodos avaliados foram: ordenação pela posição da sentença em seu texto fonte (OP), ordenação pelo tamanho da sentença em ordem crescente (OTC) e ordenação pelo tamanho da sentença em ordem decrescente (OTD). Os resultados gerados por esses algoritmos foram comparados com os sumários automáticos ordenados manualmente, do corpus CSTNews, citado acima. Os testes foram realizados com todos os 50 sumários contidos no corpus. Os resultados estão na Tabela 1.

Tabela 1: Resultados da avaliação dos métodos desenvolvidos

	Spearman		**Kendall**	
Método	**Média**	**Desvio padrão**	**Média**	**Desvio padrão**
OP	0,691	0,278	0,582	0,284
OTC	-0,186	0,558	-0,135	0,482
OTD	0,186	0,558	0,135	0,482

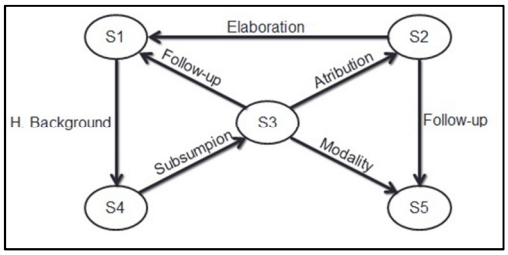
Com os resultados, podemos ver que o melhor método desenvolvido foi o da ordenação pela posição da sentença em seu texto fonte (OP). Contudo, os três métodos não nos garantem resultados aceitáveis, pois, além de seus resultados baixos, obtiveram um desvio padrão muito alto. Para melhorar esses resultados, é necessária uma abordagem mais profunda do problema, e a hipótese deste trabalho, é que, utilizando uma análise com base na teoria CST, podemos melhorar a ordenação do sumário final. Os dois métodos estudados a partir da análise da teoria CST estão a seguir.

A partir da informação semântica entre as sentenças dos textos fonte, são anotadas as relações CST. Cada uma dessas relações conectam duas sentenças, indicando-nos o tipo de informação que temos sobre esse par de sentenças, e também impondo uma direcionalidade a essa informação, caso seja necessária. Todas essas relações CST de um determinado conjunto de textos formam um grafo (G) conectando as sentenças de todos os documentos. Por exemplo, entre as sentenças "Após ter viajado para a Áustria quinta-feira, Mr.Green retornou para casa em Nova York" e "Mr.Green irá para a Áustria quinta-feira" (referenciadas por sentenças 1 e 2, respectivamente) temos uma relação "Follow-up" com direcionalidade de 1 para 2, ou seja, a sentença 1 apresenta um evento que aconteceu depois do evento narrado na sentença 2.

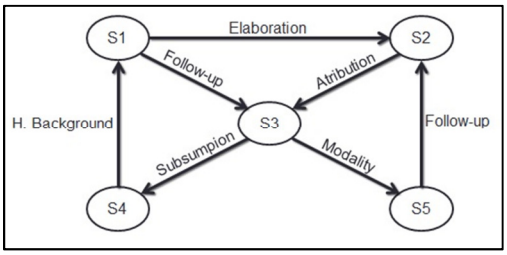
Como neste trabalho estamos interessados na ordenação entre as sentenças com base nas relações CST, é necessário verificar a informação de ordenação que existe entre as sentenças de cada relação CST, ou seja, qual sentença deve aparecer antes no sumário. Com isso, deve-se diferenciar a direção original da relação CST de sua direcionalidade lógica, que nos indica qual sentença deve aparecer antes no sumário.

Partindo desse grafo G das relações CST, mapeamos um novo grafo G' que nos indica a direcionalidade lógica da relação. Por exemplo, na relação "Follow-up" descrita acima, vemos que sua direção física se dá a partir da sentença 1 para a sentença 2, porém, podemos verificar que a melhor

ordenação entre elas, ou seja, sua direção lógica, se dá a partir da sentença 2 para a sentença 1. Os grafos hipotéticos 1 (G) e 2 (G') ilustram o problema de ordenação. Nos grafos, as sentenças provêm de diferentes textos.



Grafo 1: Exemplo de um grafo de relações CST



Grafo 2: Exemplo de um grafo de direcionalidades

Ordenando topologicamente o grafo G', obtemos todas as restrições de ordenação que as sentenças devem seguir, ou seja, obtemos uma lista onde todas as sentenças respeitam as direcionalidades de ordenação de suas relações CST. Uma ordenação possível para G' seria a sequencia seguinte de sentenças: S1, S3, S5, S4 e S2.

Como é sabido, a ordenação topológica só é possível em grafos acíclicos, sendo assim é necessária uma reestruturação do grafo G' para se aplicar a ordenação. Partindo desse princípio, temos duas abordagens para a ordenação. A primeira, mais simples, realiza a ordenação topológica sobre G', levando em conta apenas a direcionalidade das relações CST, sem se importar com o tipo da relação CST em si, excluindo os ciclos do grafo apenas com a informação de posição de cada sentença (exclui-se a sentença posicionada mais distante do início em seu texto fonte).

O segundo método avalia a semântica de cada relação CST, ponderando as arestas de acordo com a força de informação que a relação CST nos dá entre a ordem relativa entre sentenças, ou seja, qual sentença deve vir antes. Com isso, as arestas a se excluir do ciclo são escolhidas a partir de seu peso e também do número de ciclos de que ela faz parte, ou seja, quanto menor for o peso de sua relação CST e de mais ciclos ela fizer parte, ela é mais visada à exclusão. Essa abordagem foi pensada com a intuição de se excluir o menor número de arestas possível, mantendo a maior quantidade de informação semântica relevante para a ordenação. Esses dois métodos estão em fase de implementação e serão avaliados futuramente, comparando-se os seus resultados com os resultados dos métodos *baseline* anteriores.

Analisando-se os métodos ingênuos implementados no começo deste projeto, pode-se concluir que se faz necessário uma abordagem mais profunda das informações dos textos origem. Acreditamos que os métodos baseados em CST nos fornecerão isso.

Agradecimentos

À FAPESP, pelo suporte a este trabalho.

Referências

Aleixo, P. e Pardo, T.A.S. (2008). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. Série de Relatórios Técnicos do ICMC-USP/São Carlos, no. 326, 12p.

Barzilay, R., Elhadad, M., Mckeown, K. (2002). Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, Vol. 17, pp. 35-55.

Jorge, M.L.C. and Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82.

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

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In the *Proceedings of the Annual Meeting of the ACL*, pp. 545-552.

Okazaki, N.; Matsuo, Y.; Ishizuka, M. (2004). Improving Chronological Sentence Ordering by Precedence Relation. In the *Proceedings of the 20th International Conference on Computational Linguistics*.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross- document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.