Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Núcleo Interinstitucional de Linguística Computacional – NILC

Criação de Sumários Extrativos e Abstrativos com base em Aspectos de Opiniões de Livros e Produtos Eletrônicos

Aluno: Roque Enrique López Condori
Orientador: Thiago Alexandre Salgueiro Pardo

1. Introdução

A sumarização de opiniões, também conhecida como sumarização de sentimentos, é a tarefa que consiste em gerar automaticamente sumários para um conjunto de opiniões sobre um alvo especifico (Conrad et al., 2009). Uma das principais abordagens para gerar sumários de opiniões é a sumarização baseada em aspectos. A sumarização baseada em aspectos produz resumos das opiniões para os principais aspectos de um objeto. Os objetos normalmente referem-se a produtos, serviços, organizações, pessoas, eventos, entre outros, e os aspectos são atributos ou componentes dos objetos (Zhang e Liu, 2013). Segundo Hu e Liu (2004), um aspecto, em termos gerais, é uma característica ou função de um produto.

Um sistema automático de sumarização de opiniões baseado em aspectos tem como entrada um conjunto de comentários sobre um objeto (por exemplo, um restaurante) e produz um resumo que expressa o sentimento para alguns aspectos relevantes (por exemplo, a comida ou o serviço do restaurante).

No desenvolvimento de métodos automáticos de geração de sumários baseados em aspectos é necessário, muitas vezes, um córpus de resumos criados manualmente por humanos para assim poder avaliar a qualidade destes métodos. Infelizmente para a língua portuguesa do Brasil ainda não existe um córpus de resumos de opiniões baseados em aspectos.

Devido a essa ausência, o objetivo principal desta anotação visa criar tanto sumários extrativos, que contenham as opiniões principais dos usuários, quanto sumários abstrativos, que sintetizem de maneira mais refinada as opiniões expressas.

Para esta anotação foi considerado o córpus ReLi (Freitas et al., 2012), um córpus de resenhas de livros no qual as opiniões já tem identificados os aspectos. Nesta tarefa, foi considerada também uma pequena amostra de comentários sobre 4 produtos eletrônicos do website Buscapé.

2. Metodologia de anotação

2.1 Considerações Iniciais

- Se estima que 14 pessoas participariam da tarefa de criação de sumários.
- Para esta tarefa se consideram 17 entidades: 13 livros do córpus ReLi e 4 produtos eletrônicos do córpus Buscapé.
- Cada entidade terá 10 comentários diferentes (Carenini et al. (2006), Tadano et al. (2010)).

- Para cada entidade pretende-se criar 5 sumários extrativos e 5 sumários abstrativos.
- Em total cada pessoa fará 12 sumários (aproximadamente 6 extrativos e 6 abstrativos).
- A duração da tarefa é 12 dias úteis, cada dia uma pessoa fará só um sumário extrativo ou abstrativo.
- A criação de sumários será feita todos os dias da semana (de segunda-feira, até sexta-feira). Aproximadamente em 2 semanas e 2 dias a anotação será concluída.
- A criação de sumários poderá ser realizada a distancia e enviada pelo e-mail com um prazo limite até o dia posterior.
- Na Tabela 1 mostra-se a quantidade de sumários extrativos e abstrativos que serão criados nesta tarefa.

	Entidades	**Sumários Extrativos**	**Sumários Abstrativos**	**Sumários em Total**
Livros (Córpus ReLi)	13	5	5	130
Produtos eletrônicos (Córpus Buscapé)	4	5	5	40
				170

Tabela 1: Quantidade de sumários extrativos e abstrativos

2.2 Descrição da Tarefa

Nesta tarefa pretende-se criar sumários extrativos e sumários abstrativos para as 17 entidades selecionadas.

A. Sumários Extrativos

- Os sumários abstrativos serão compostos por 100 palavras aproximadamente (Carenini et al. (2006)), com 10 palavras para mais ou para menos.
- Não será permitida a reescrita das sentenças dos textos originais. Se as sentenças presentam erros ortográficos, gramaticais, etc., esses não devem ser corregidos.
- Os sumários gerados devem abranger os principais aspectos presentes nos comentários (revisar Tabela 2).
- Não considerar sentenças sobre os filmes dos livros no sumário final.
- Os tags presentes no final de cada sentença não devem ser removidos (exemplo: <D7_S1>). Por outro lado, esses tags não devem ser contados como palavras no cálculo do tamanho do sumário (100 palavras aproximadamente).

B. Sumários abstrativos

- Os sumários abstrativos serão compostos por 100 palavras aproximadamente (Carenini et al. (2006)), com 10 palavras para mais ou para menos, e devem ser o mais reescrito possível.
- Os sumários gerados devem abranger os principais aspectos presentes nos comentários (revisar Tabela 2).
- Os sumários abstrativos gerados devem mostrar o sentimento (positivo ou negativo) sobre as entidades e os aspectos.
- Não considerar informações sobre os filmes dos livros no sumário final.

Entidades	**Aspectos**
Livro (ReLi)	leitura, personagens, enredo, historia, trama, linguagem, texto, capítulos, romance, narrativa, diálogos, tema, escrita, frases, protagonista, título, imagens, início do livro, metade de o livro, fim do livro, vocabulário.
Smartphone (Buscapé)	bateria, design, processador, tela, preço, câmera, peso, sistema operacional, internet, foto, vídeo, wi-fi, som, tamanho, fones de ouvido, velocidade, chip
TV (Buscapé)	design, preço, câmera, qualidade da imagem, luminosidade, wi-fi, som, durabilidade, internet

Tabela 2: Entidades y aspectos considerados na anotação

2.3 Material de Trabalho

- Cada participante da tarefa de criação de sumários deverá utilizar seu computador pessoal ou notebook.
- As 10 opiniões de cada produto serão fornecidas em um único arquivo via Dropbox no formato *.txt*. (https://www.dropbox.com/sh/npjywqmu12ks3oy/AABVYLh9vyovmjZHI7hjtjL4a?dl=0)
- Os sumários gerados serão armazenados em arquivos *.txt* em codificação UTF-8 (revisar Figura 1), considerando o seguinte formato para os nomes dos arquivos:

  "<Livro/Produto>_extrativo_NomeCriadorDoSumário.txt" e
  "<Livro/Produto>_abstrativo_NomeCriadorDoSumário.txt"

- Opcionalmente, os participantes podem enviar, em um arquivo separado, comentários sobre o processo de criação de um determinado sumário. Nos comentários, os participantes podem destacar alguma informação que julguem importante, por exemplo: dificuldades, sentimento percebido nos comentários, etc.

- O envio dos sumários criados será feita pelo e-mail ao endereço rlopezc27@gmail.com

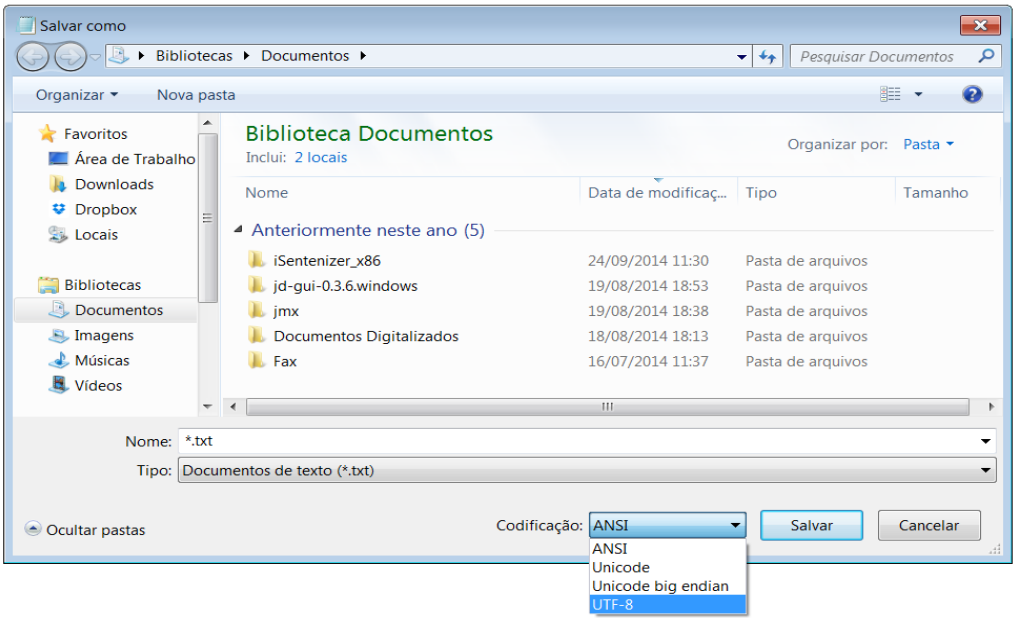

Figura 1: Salvar arquivo em codificação UTF-8 em Windows

Referências Bibliográficas

Carenini, G.; Ng, R.; Pauls, A. (2006). Multi-document summarization of evaluative text. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pp. 305_312.

Conrad, J. G.; Leidner, J. L.; Schilder, F.; Kondadadi, R. (2009). Query-based opinion summarization for legal blog entries. In ICAIL, pp. 167-176.

Freitas, C.; Motta, E.; Milidiú, R.; Cesar, J. (2012). Vampiro que brilha... rá! desaos na anotação de opinião em um corpus de resenha de livros. In Proceedings do XI Encontro de Linguística de Corpus.

Hu, M.; Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, New York, NY, USA, pp. 168-177.

Tadano, R.; Shimada, K.; Endo, T. (2010). Multi-aspects review summarization based on identification of important opinions and their similarity. In PACLIC, pp. 685_692. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Wang, D.; Zhu, S.; Li, T. (2013). Sumview: A web-based engine for summarizing product reviews and customer opinions. Volume 40, Tarrytown, NY, USA, pp. 27_33. Pergamon Press, Inc.

Zhang, L.; Liu, B. (2013). Aspect and entity extraction for opinion mining. In W. W. Chu (Ed.), Data Mining and Knowledge Discovery for Big Data, Volume 1 of Studies in Big Data, pp. 1-40.

Zhu, J.; Zhu, M.; Wang, H.; Tsou, B. K. (2009). Aspect-based sentence segmentation for sentiment summarization. In Proceedings of the 1st International CIKM workshop on Topic-sentiment analysis for mass opinion, TSA '09, New York, NY, USA, pp. 65-72.