Quem, o Que, Onde

Lucía Castro & Lucia Rino

Who wrote What Where

- Owczarzak, Karolina; Dang, Hoa Trang (2011).
  - Who wrote What Where: Analyzing the content of human and automatic summaries. *Proc. of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pp. 25–32. Portland, Oregon, June 23. Association for Computational Linguistics.

Quem, o Que, Onde

- Foco

Foco do artigo

- Abstração vs. extração na SA
- Genest et al. (2009)
  - Sumarizadores automáticos "top" comparáveis a sumários extrativos manuais
    - Mas menos "responsivos", ou **menos adequados à pergunta**, do que os abstracts manuais
      - **Adequação**
- Autores pensam havermos chegado ao limite das técnicas extrativas
  - Necessidade de técnicas de abstração

"Parênteses"

- **Responsiveness** (DUC'2003)
- Task 1 – Very short summaries
- Task 2 – Short summaries focused by events
- Task 3 – Short summaries focused by viewpoints
- Task 4 – Short summaries in response to a question
  - Contexto de Q&A
  - Sumários curtos para responder a uma consulta/pergunta
  - "Responsivos" na forma e no conteúdo
- Quão bem um sumário responde a uma consulta/pergunta
  - Graus: 0 (pior), 1, 2, 3, or 4 (melhor)

Quem, o Que, Onde

- O problema

SA profunda

- Abstração (K. Sparck Jones, 1993)
  - Seleção do que é importante
  - Exclusão do que é irrelevante
  - Redução da info
    - Generalização do particular e específico
    - Parafraseamento
    - Identificação de estruturais gerais ou globais

SA profunda

- Abstração (K. Sparck Jones, 1993)
  - O que qualifica uma info importante/saliente/relevante?
    - **O que é** informação relevante num dado cenário?
  - Como achamos essa info de interesse ?
    - **Como** encontrá-la?
    - **Quando** um sumário é melhor que outro?

⇨ Respostas subjetivas, portanto, não definitivas

Proposta ⇨ Sumarização Guiada

TAC 2010

- Guided Summarization

TAC 2010

- **Summarization**
  Track Coordinators: Karolina Owczarzak (karolina.owczarzak@nist.gov) and Hoa Trang Dang (hoa.dang@nist.gov)
  Web page: http://www.nist.gov/tac/2010/Summarization/
  Mailing list: duc_list@nist.gov

TAC 2010

- Não falam em *resolver*
  - O que qualifica uma info importante/saliente/relevante ?
  - Como achamos essa info de interesse ?
- Falam em *neutralizar esses problemas*
  - Categorias topicais
  - Listas de aspectos que um sumário adequado deve contemplar

TAC 2010

- Objetivos

TAC 2010

- Para *neutralizar*
  - *O QUE*
  - *COMO/ONDE*

- Busca-se similaridade com modelos humanos
- Alvo bem definido: SA guiada
- "Diagnostica-se" o conteúdo dos sumários

- E, assim, buscam-se sumarizadores automáticos orientados pelo significado

SA guiada

- OBJ: encorajar uma análise linguística mais profunda (semântica) dos TFs
- Sumário de ~100 palavras
- Multidoc AS: 10 artigos jornalísticos de um certo tópico

- Tópicos categorizados previamente
  - Categorias topicais

SA guiada

- Cada categoria
- Caracterizada por uma lista de aspectos
- Sumário deve incluir todos os aspectos para sua categoria

SA guiada

- Cada categoria
- Caracterizada por uma lista de aspectos
- Sumário deve incluir todos os aspectos para sua categoria

TAC 2010

- Hipóteses

Hipóteses para tópicos categorizados

- Estórias do mesmo tipo seguem um *template* específico
- Incluem o mesmo tipo de fatos

- Então, podemos determinar:
  - **O QUE** é importante
  - **ONDE** procurar

Categorias

1. Accidents and Natural Disaster
2. Attacks
3. Health and Safety
4. Endangered Resources
5. Trials and Investigations

Sucinto/2012 20

Sumarização Guiada

- *Guided Sumarization* – TAC 2010
  - Categorias→ Aspectos
  - p.ex.

Accidents	Attacks	Health
what	what	what
when	when	who affected
where	where	how
why	perpertrators	why
who affected	why	countermeasures
damages	who affected	
countermeasures	damages	
	countermeasures	

Aspectos semânticos sugeridos

Aspectos	Casos (semânticos) sugeridos
what happened	O QUE
who is **under** investigation	QUEM (paciente)
who is investigat**ing/**suing	QUEM (agente)
who is **affected**	QUEM (vítima)
date	QUANDO
why (general)	CAUSA
description of resource	DESCRITIVO
location	LOCATIVO
how they are affected	COMO (algo é afetado)
why it happens	PORQUE (causa/justificativa)
threats to resource	DANOS

Aspectos semânticos sugeridos

OUTROS
specific charges
importance of resource
casualties
countermeasures
sentence/consequences
damages
rescue efforts/countermeasures
how do they plead/react to charges
perpetrators

TAC *Guided Sumarization Task 2010*

- 5 Categorias
  - acidentes e desastres, ataques, saúde e segurança, juízos e recursos
  - TAC forneceu aspectos
  - 20 textos para cada categoria
  - 4 sumários modelo/categoria
- Deviam-se produzir dois tipos de sumários:
  - Iniciais→ 10 primeiros textos
  - Atualizados→ 10 seguintes textos *tratamento de Redundância*

TAC 2010

- Avaliação

Sumarização Guiada

- Objetivos:
  - tarefa de sumarização mais focada (maior concordância)
  - Utilização de métodos mais linguísticos
  - ferramenta diagnóstico para análise de conteúdo
- Owczarzak e Trang (2011) mostram como esses objetivos foram alcançados na TAC 2010
  - Avaliação → Qualidade, Conteúdo (Método da Pirâmide) (Nenkova e Passoneau, 2004) e *Responsiveness*

Método tradicional de avaliação de sumários da DUC (TAC) e o Método da Pirámide

Nenkova e Passoneau (2004)

Métodos da DUC 2003 e anteriores

- Humano cria sumário modelo de 100 palavras
- Modelo é segmentado
  - Edu's
- sumário automático é segmentado :
  - Sentenças, Edu´s, etc.
- Evalúa-se:
  - Pares de unidades com overlap de informação
  - % de conteúdo no overlap

Desvantagens dos métodos tradicionais

- Resultados pouco confiáveis
  - Apenas um único sumário modelo
- Pouca concordância entre apenas dois sumários
  - Pontuações sempre baixas

Método da Pirámide

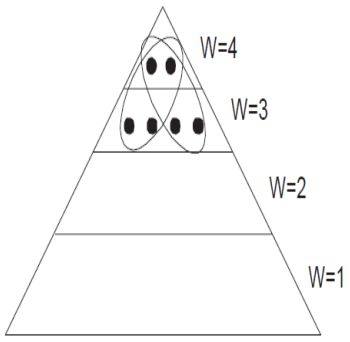
- Método baseado na análise de Unidades de Conteúdo do Sumário (SCUs)
- Exemplos de SCU

A1 In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya.
B1 Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988.
C1 Two Libyans, accused by the United States and Britain of bombing a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain.
D2 Two Libyan suspects were indicted in 1991.

Método da Pirâmide

- Estágios do processo
  - identificação de Sentenças similares
  - Indentificação de SCUs e contribuidores (unidades que compõem SCU→ definem contexto semântico)
  - pontuação de SCUs
    - # de sumários nos quais aparecem
  - SCUs são particionadas numa pirâmide de acordo com a pontuação

Método da Pirâmide



- W é o numero de sumários modelo para formação da pirâmide
- Sumários Automáticos com conteúdo mais no topo da pirâmide → mais *ótimos*

Método da Pirâmide

- Como computar a pontuação de uma SCU de acordo com os dados da pirâmide

$\mathcal{D} = \sum_{i=1}^{n} i \times D_i$ '/Di = # SCU do sumário no nível i da pirâmide

TAC 2010

- Resultados da aplicação do método da pirâmide

TAC *Guided Sumarization Task 2010*

- Aplicação do método da pirâmide
  - Sumários iniciais e atualizados
- Resultados mostraram que:
  - pontuações de sumários automáticos
    - Valores maiores de W na pirâmide
    - Diferencia acentuada entre sumários iniciais e atualizados
  - em sumários manuais
    - Distribuição usual de pirâmides para categorias: Saúde e Juízos
    - Em categorias como ataques e acidentes acontece diferente

Resultados TAC *2010*

- Sumários automáticos
  - Melhores pontuações em categorias do tipo Ataques e Acidentes
- Essas categorias têm aspectos que tendem a gerar as mesmas respostas na maioria dos casos manuais
  - Ataques → *quando, onde*
  - Acidentes → *o que, quando, onde, perpetrators, who_affected*
- Porque mais previsíveis, segundo os autores

Resultados TAC *2010*

- Métodos de seleção de conteúdo podem identificar alguns dos fatos mais importantes
- A densidade da informação é menor nos sumários automáticos do que nos manuais
  - Conteúdo com compressão inadequada
  - Conteúdo irrelevante
  - Conteúdo redundante

Resultados TAC *2010*

- Sumarizadores automáticos "top"
- A maioria baseada na posição da sentença
  - Mas posição pode ser um indicador melhor só para alguns tipos de informação
- Melhores resultados para *eventos*
  - Ataques, acidentes, julgamentos
  - Grande parte das notícias

Sugestões

- Para o futuro

Questões para pensar

- Aplicar SA guiada ao português?
- Como?
- Método da Pirâmide para análise multidocumento?
- Simulação do método da Pirâmide

Questões para pensar

- Genest (2009)
  - Extratos humanos comparados com abstracts humanos
    - ROUGE
      - Extratos humanos piores que abstracts?
- 2 formas de fazer
  1. Considerar tamanhos reais (diferentes), sem truncar
  2. Truncar para comparar
     - Para truncar, aplicar a ideia do desvio da taxa de compressão: pegar a melhor aproximação

- CSTSumm (Castro)?
  - Atingiu o limite pelo método extrativo?
  - Próxima reunião
  - 13/04 – quali Verônica
  - Lucía – experimento
    - 27/04 – tarefa
    - TODOS – ler Genest (2009)
    - Estudo sobre amostras do Corpus

Proposta Thiago

- CSTNews (p/ 27/04)
- Já categorizados
- Procurar aspectos
  - Começar pelos aspectos do artigo
    - Empiricamente
    - p/ aspectos inexistentes, criar subjetivamente
- Expandir (após dia 27/04)
  - Manual da folha
  - Estudos de gênero (Ani)
- Retextualização = reescrita/abstração (Sparck Jones)

Proposta Lucia

- Outros artigos sobre o Método da Pirâmide
- TAC 2012
- Usar o Met Piram para sumarizar
  - Alguém fez isso?