Detecção de Expressões Temporais para Sumarização Multidocumento

Luiz Antonio de Menezes Filho e Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguísticas Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

Luiz.menezesf@gmail.com, taspardo@icmc.usp.br

Abstract. *This paper presents a proposal for temporal expression detection in news texts for multi-document summarization.*

Resumo. *Apresenta-se, neste artigo, uma proposta para detecção de expressões temporais em textos jornalísticos para fins de sumarização multidocumento.*

1. Introdução

O processamento de tempo em textos, tarefa relativa à área Processamento de Línguas Naturais (PLN) tem obtido cada vez mais interesse do mundo científico, devido a seus significantes benefícios às tarefas multidocumento, como a sumarização automática. Por exemplo, para a sumarização, o processamento temporal pode auxiliar na ordenação cronológica dos eventos narrados nos sumários, provendo assim, um sumário mais coerente e coeso.

A resolução de expressões temporais (ET) consiste em identificar, classificar e obter a data de expressões temporais (expressões como "12 de janeiro" ou "ontem") de um texto. Por exemplo, a palavra "amanhã" seria identificada como ET, classificada e, por fim, seria obtido um valor normalizado (uma representação da data facilmente interpretada por computador) para ela que indique, a partir da data em que o texto foi feito, a data que tal expressão referencia.

Este trabalho propõe um método para resolução de expressões temporais com a utilização de expressões regulares (*regex*), as quais foram criadas a partir do conhecimento proveniente de um córpus, e demonstra os resultados obtidos por um sistema implementado.

A Seção 2 apresenta os trabalhos que influenciaram os métodos de resolução de expressões temporais deste trabalho. A Seção 3 apresenta o método em si e os resultados obtidos.

2. Trabalhos Relacionados

A *Message Understand Conference* definiu o TIMEX, um esquema para classificar expressões temporais. Mani e Wilson (2000) também apresentaram um possível esquema de classificação de ETs e técnicas para anotação dessas expressões. O TIDES (*Translingual Information Detection, Extraction and Summarization*) (Ferro et al., 2001) introduziu o TIMEX2, cuja estrutura possibilita uma melhor classificação de ETs. Outro trabalho nessa linha é o TimeML (Saurí et al., 2005), que apresenta o TIMEX3.

A pesquisa que mais influenciou os métodos utilizados por este trabalho foi o Segundo HAREM (Mota e Santos, 2008), que consistiu em uma avaliação conjunta (ou seja, vários sistemas concordam em fazer uma tarefa, a fim de comparar o desempenho entre eles) relacionada a entidades mencionadas em português. Dentre as entidades mencionadas abordadas no Segundo HAREM, há a categoria TEMPO, descrita por (Baptista et al., 2008). Este último fornece diretrizes para identificação, classificação e obtenção da data que as expressões temporais referenciam (nesse trabalho esta tarefa foi denominada "normalização de ETs").

3. Método de Detecção de Expressões Temporais

O córpus utilizado neste trabalho foi o *CSTNews* (Aleixo e Pardo, 2008), que contém aproximadamente 140 notícias retiradas de jornais como Jornal do Brasil, Gazeta do Povo, O Globo e outros. O córpus foi anotado segundo as diretrizes propostas por (Baptista et al., 2008) e essas diretrizes guiaram a proposta de expressões regulares. Foram anotadas cerca de 1000 expressões temporais. A marcação segue o padrão XML e a quantidade de atributos de cada

marcador varia de acordo com o tipo de cada expressão. Para melhor entendimento, segue um exemplo de ET anotada do córpus *CSTNews.*

> *"O último jogo havia sido* ***<ET ID="03" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="200510--T----E--LM-">*** *em outubro de 2005* ***</ET>****, na vitória por 3 a 0 sobre a Venezuela (...)."*

A classificação das expressões temporais proposta por Baptista et al. é mostrada na Figura 1.

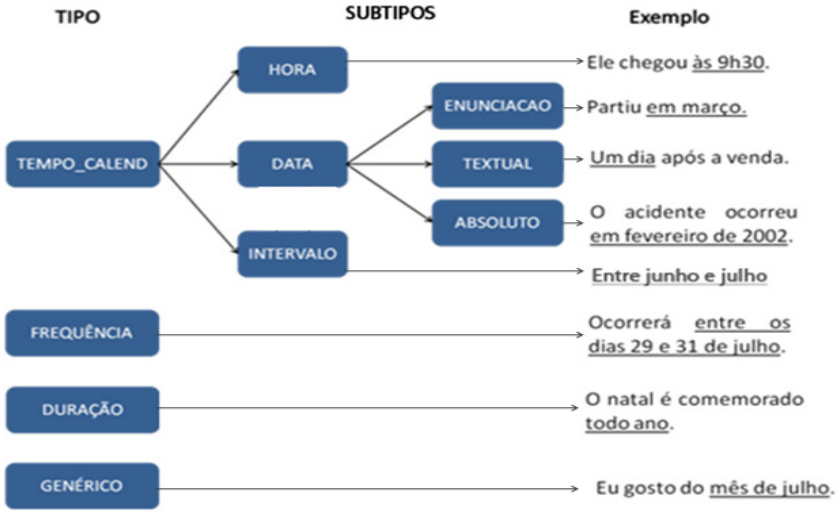

Figura 1. Esquema da classificação de Expressões Temporais

Para melhor modelagem da identificação das expressões temporais, estas foram divididas em oito grupos: Absoluto, Duração, Freqüência, Intervalo, Hora, Enunciação, Textual, Genérico. Também foram contabilizados (i) expressões temporais que foram anotadas, mas não se chegou a um consenso sobre sua classificação, e (ii) os "cabeçalhos", que são ETs que indicam a data em que o texto foi criado ou publicado. A Tabela 1 contém a contagem da ocorrência de expressões temporais no córpus *CSTNews* divididas nesses grupos.

Tabela 1. Ocorrência de expressões temporais no córpus *CSTNews*

Tipo de ET	Ocorrência	Tipo de ET	Ocorrência
Absoluto	121	Enunciação	522
Duração	54	Textual	28
Frequência	6	Genérico	1
Intervalo	76	Sem classificação	3
Hora	122	Cabeçalhos	140

Para facilitar a implementação e obter resultados mais rapidamente, a tarefa de resolução de expressões temporais foi dividida em duas fases, uma de identificação e outra de resolução (que consiste em classificações não triviais e normalização das ETs).

Na etapa de identificação, foram utilizadas regex para identificar as expressões temporais, pois a maioria das ETs apresentam padrões que podem ser previstos. Seguem abaixo alguns exemplos dessa abordagem, sendo as partes sublinhadas da sentença o padrão buscado e, abaixo, a regra proposta para o sistema.

O acidente ocorreu no dia 23 de agosto de 1859.

(Nn)o" dia "{dia}" de "{mes}" de "{ano}

Para identificar automaticamente as expressões temporais, foi utilizado um gerador de analisador léxico, o LEX. Por esse motivo, as expressões regulares criadas seguem o padrão desse programa. As palavras entre colchetes ("{}") indicam uma redefinição, ou seja, há outra regra, fora da regra em que aparece essa redefinição, que define esta. Por exemplo, {dia} é definido como a seguir: dia [0-9]([0-9]?)(º)?

Até o momento foram implementados os identificadores dos grupos: absoluto, duração (em fase de revisão) e o de freqüência. Cada um identifica seus respectivos tipos de expressões temporais e já faz uma classificação superficial.

A anotação feita automaticamente foi avaliada pela comparação dos resultados da anotação automática com a manual (que foi considerada como a anotação ideal). Foi levantada a quantidade de anotações incorretas do sistema.

A Tabela 2 mostra os resultados obtidos até o momento. Por exemplo, a segunda linha dessa tabela mostra informações relativas às expressões temporais do tipo Absoluto. A segunda coluna da tabela demonstra a ocorrência de expressões temporais do tipo Absoluto no córpus CSTNews. Da terceira à sexta coluna, são mostradas as anotações incorretas feita pelo sistema. Essas anotações incorretas foram divididas em quatro: Adicional (ETs que foram anotadas pelo sistema, mas não existem no córpus), Faltante (ETs existentes no córpus que não foram anotadas pelo sistema), Diferente (ETs identificadas corretamente, mas classificadas erroneamente) e Total. A última coluna contabiliza as expressões anotadas corretamente.

Tabela 2. Resultados obtidos pelo sistema de resolução de expressões temporais

		Incorretas				
Tipo de ET	Córpus	Adicional	Faltante	Diferente	Total	Corretas
Absoluto	121	26	1	3	30	91
Duração	54	22	34	4	60	54
Frequência	6	20	0	1	21	6

Pelos resultados obtidos, observou-se a necessidade de reformular as expressões regulares que identificam os tipos duração e frequência. Quanto à freqüência, pode-se dizer que esse resultado se dá pela falta de exemplos no córpus, pois os padrões foram criados a partir dessas ocorrências no *CSTNews*. Também se leva em conta uma revisão do córpus a fim de verificar se essa baixa quantidade de ETs do tipo freqüência não foi gerada por problemas da anotação manual.

Os melhores resultados obtidos foram das ETs do tipo Absoluto, que tiveram uma taxa de acerto de aproximadamente 75%, sendo que a maioria dos erros veio de marcações excedentes. Estas foram analisadas mais a fundo e chegou-se a conclusão de que boa parte dessas marcações excedentes ocorria porque algumas ETs do tipo intervalo eram erroneamente marcadas como absolutas. Assim, esses erros podem ser removidos quando o módulo que identifica as expressões temporais de intervalo for implementado.

As expressões do grupo genérico só aparecem uma vez no córpus. Uma possível explicação para esse acontecimento é que notícias (tipo de texto do córpus *CSTNews*) não utilizam muitas expressões temporais genéricas. Quanto à parte de identificação, ainda não foram feitos os módulos para intervalo, hora, enunciação e textual. Depois que esses forem feitos e avaliados, começará a ser feita a parte de normalização das expressões temporais.

Agradecimentos

Ao PIBIC/CNPq e 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 CST (Cross-Document Structure Theory)*. Série de Relatórios Técnicos do ICMC-USP, no. 326.

Baptista, J.; Hagège, C.; Mamede, N. (2008). Capítulo 2: Identificação, classificação e normalização de expressões temporais do português: A experiência do segundo HAREM e o futuro. Em C. Mota e D. Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Ferro, L.; Mani, I.; Sundheim, B.; Wilson, G. (2001). *TIDES Temporal Annotation Guidelines*. MITRE Technical Report, MTR 01W0000041.

Mani, I. and Wilson, G. (2000). Robust Temporal Processing of News. In the *Proceedings of the ACL Conference*.

Mota, C. e Santos, D. (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Saurí, R.; Littman, J.; Knippen B.; Gaiazukas, R.; Setzer, A.; Pustejovsky J. (2006). *TimeML Annotation Guidelines Version 1.2.1*.