Reconhecimento e normalização de expressões temporais no HAREM 2

C. Hagège
Xerox Research Centre Europe, Grenoble (France)

J. Baptista e N. Mamede
L2f INESC-ID Lisboa (Portugal)

Plano da apresentação

- Motivações da proposta
- Dificuldades de caracterização
- Proposta para o HAREM
- Balanço e próximos passos

Motivações da proposta

Motivações

- Necessidade de tomar em consideração a vertente temporal em tarefas de pesquisa de informação

  Qual é a capital de Alemanha ?

  Quem é o Presidente de Portugal ?

- Trabalho recente na área
  - TempEval

    TimeML e TimeBank (*TimeML Annotation Guidelines Version 1.2.*, Roger Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, James Pustejovsky)
  - Google Time

META: localizar / estruturar / ordenar temporalmente os eventos descritos num texto.

Algumas dificuldades na caracterização e representação

Dificuldades na caracterização e representação (1)

- Ambiguidade dos sinais introdutores das expressões temporais (ET)

  em 2008

  em duas semanas

- Interpretação vaga

  Como interpretar : *Há dois anos ?*

Dificuldades na caracterização e representação (2)

- Necessidade de tomar em consideração elementos externos para cálculo da própria ET

  O avião vai aterrar no domingo.

  O avião aterrou no domingo.

- Dificuldade de encontrar uma representação coerente para certas expressões temporais

  e.g. Frequências e Agregados temporais

  Todas as primeiras quartas-feiras do mês

  Quatro domingos consecutivos

Proposta para o HAREM

A proposta para o HAREM 2

- Restrições
- Caracterização das ET
- Tokenização das ET complexas
- Normalização

Restrições

Uma tarefa executável em 6 meses de desenvolvimento

- → Compatibilidade com propostas já existentes
- → Tentar limitar a dependência entre eventos e ET
- → Tornar a tarefa TEMPO independente da problemática da subcategorização verbal (inclusão de qualquer preposição na expressão temporal)
- → Definir critérios claros de tokenização das ET
- → Classificar *antes* de resolver a referência temporal
- → Normalização parcial das ET
- → Agregados temporais não considerados na sua especifidade
- → Tentar assegurar ao máximo o critério de intersubjectividade na anotação (listagem e descrição dos elementos lexicais que entrem na formação das ET)

Caracterização das ET

- **Expressões calendárias**
  - **Datas**
    - Absolutas *10 de Novembro de 2008*
    - Relativas (ou referenciais) *ontem, no dia anterior*
  - **Horas** *às 5h30*
  - **Intervalos** *entre 3 e 15 de Janeiro*
- **Durações** *durante dois anos*
- **Frequências** *dois dias por semana*
- **ET genéricas** *A Primavera* é a mais bela estação do ano

Datas absolutas

<CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" REF="ABSOLUTO"/>

Ideia subjacente

- não necessitam de informação externa
  para poderem ser colocadas na linha temporal.
- são sujeitas a **normalização**.

  Exemplo *15 de Janeiro de 2007*

Limitação

Qual é a granularidade que se pretende ? O dia, o mês, o ano ?
Que fazer com *2003* ? F*evereiro* ? *dia 2* ? *Primavera ?*

Por convenção

Decidimos que, se pelo menos um dos campos ANO, MÊS, DIA ou ESTAÇÃO estiver preenchido, se trata de uma data absoluta.

Datas referenciais (1)

Ideia subjacente

As datas referenciais correspondem a ET calendarizáveis mas que só mediante cálculo a partir de outra data de referência.

Dois tipos de referência

Referência ao momento da enunciação

A calendarização da ET é calculada a partir do momento em que for proferida a asserção ou à data de criação do documento (DCT);

i.e. o contexto discursivo não é necessário para interpretar a ET

Referência textual

A calendarização da ET está relacionada com a data de ocorrência de um outro evento ou de outra ET presente no texto.

i.e. sem o contexto discursivo é impossível interpretar a ET

Datas referenciais (2)

Exemplos

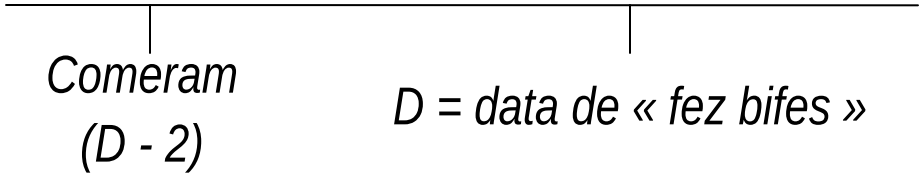
REF=ENUNCIACAO

Há dois dias comeram um belo bacalhau.

DCT/Momento de Enunciação = 7 de Setembro de 2008

REF=TEXTUAL

Fez uns bifes porque *dois dias antes* comeram um belo bacalhau



Datas referenciais (3)

No âmbito do HAREM 2

- distinguir os dois tipos de datas referencias (enunciação ou textual)
- pistas para a normalização

  i.e., quando for possível:

  - completar o valor do atributo SENTIDO que indica se a data referencial é anterior, posterior ou simultânea à data da referência (seja qual for o tipo de referência)
  - completar o valor do atributo VAL_DELTA que indica a distância temporal entre a referência e a expressão referencial.

Exemplo

Apareceu <TEMPO TIPO=”TEMPO_CALEND” SUBTIPO= "DATA" REF=“TEXTUAL" **SENTIDO=“POSTERIOR" VAL_DELTA="A0M0S2D0H0M0S0">** *duas semanas* </TEMPO> **depois d***a festa*.

Datas referenciais (4)

Limitação

Certa arbitrariedade devida à convenção que adoptamos para classificar datas absolutas

Exemplo *no dia 2*

Simplificação do momento de referência (enunciação ou textual)

Exemplo

*O Pedro repondeu **na semana passada**: "Cheguei **ontem**".*

na semana passada

⟶ D(*Pedro responder*)ref:enunciação = DCT – 1 semana

ontem

⟶ D(*Pedro chegar*)ref:enunciação = D(*Pedro chegar*) – 1 dia

Horas

<CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" />

- Expressões de DATA absolutas mas com granularidade inferior ao dia.
- São sujeitas a normalização

Exemplo *O Pedro está disponível às 14:00 horas*

Limitação

- Distinguimos por convenção o subtipo HORA do subtipo DATA absoluta, mas tal não é absolutamente essencial (sendo possível unificar a representação de ambos os subtipos)
- Além disso, podem existir horas referenciais (eg. *uma hora depois*, *várias horas mais tarde*) que funcionam exactamente do mesmo modo que outras datas referenciais

Intervalos

<CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" />

- Expressões temporais complexas
- denotam um intervalo de tempo com **dois limites explícitos**
- semanticamente, podem ser considerados como datas (localizam temporalmente o evento)
- não são, por agora, normalizados

Exemplos

Trabalhei em Londres ***entre 2000 e 2003.***
Trabalhei em Londres ***de Outubro a Dezembro de 2007.***

Limitação

O critério de "dois limites explícitos" não é suficiente para assegurar que a expressão é um intervalo
eg. *Vai demorar* ***entre 3 e 6 dias*** (duração complexa)

Durações

- Expressões temporais que denotam uma duração de tempo contínuo.
- Ao contrário das datas, não exprimem a localização temporal do evento que modificam mas sim uma quantificação temporal.
- Respondem à interrogativa *(prep) quanto tempo ?*

Exemplos

Fiquei **dois meses** em Lisboa.

O urso fica **todo o inverno** na toca.

Frequências

- Expressões temporais que denotam a repetição de um evento no tempo.

Exemplos

*Vou ver os meus pais **diariamente.***
*Vou ver os meus pais **todos os dias.***
*Vou ver os meus pais **duas vezes por semana.***
*Vou ver os meus pais **dia sim dia não.***

ET Genéricas

- Expressões constituídas por elementos lexicais que também entram em expressões temporais.
- MAS:
  - que **não permitem calendarização** de eventos associados (datas, intervalos),
  - **não denotam duração** (durações),
  - **nem repetição** de eventos ao longo da linha do tempo (frequências).

Exemplos

Adoro o verão.

Fevereiro é o mês mais curto do ano.

Critérios para a tokenização de ET complexas

Considerando uma ET complexa e o evento (EVT) por ela modificado temporalmente, esta ET complexa deve ser subdividida em sub-expressões (SExp) sse:

1 - cada SExp é sintácticamente válida quando ligada ao EVT e
2 - cada uma das combinações EVT + SExp é logicamente implicada pela combinação EVT + ET

Exemplos

Falamos [*duas vezes por semana*] (1)
Viajamos [*dez dias*] [*em Setembro*] (2)
Aconteceu um dia antes do Natal. (Ambíguo)

Normalização

Tarefa de normalizacao limitada

- Tentativa de normalizar *totalmente :*
  - Datas absolutas
  - Horas
- Normalização preliminar de datas referenciais, através dos atributos SENTIDO e VAL_NORM
- Nada determinado para as outras sub-categorias

Balanço da tarefa e próximos passos

Participação

- **7** sistemas participaram na tarefa TEMPO
- diferentes graus de participação :
  - 6 sistemas com TIPO
  - 5 sistemas com SUBTIPOS
  - 2 sistemas com REF
    (tipo de referência para datas referenciais)
  - 1 sistema com a normalização

→ **prudência** e **moderação** no desenvolvimento da tarefa para futuras avaliações conjuntas de TEMPO
(o que não impede melhoramentos/correcções)

(Ver resultados na página do HAREM@Linguateca)

Próximos passos

Num futuro breve

- Não fazer a distinção entre datas e horas.
- Ser mais restritivos no que consideramos como data absoluta (alargar a noção de data referencial)
- Aperfeiçoar algumas definições (intervalos vs. duração, listagem exaustiva dos constituintes das ET).

Próximos passos

Num futuro menos breve

- Começar o cálculo e a normalização de algumas datas referenciais (as ET referenciais a DCT)
- Especificar com mais precisão para as datas referenciais a expressão linguística que lhes serve de referência
- Considerar a *vagueza* de certas expressões temporais na sua normalização
- Reflectir sobre uma representação adequada dos agregados temporais (TimeML ?)

E depois ?

Ordenação temporal (parcial) dos eventos relatados nos textos (inferência temporal).

Obrigado !

C. Hagège
Xerox Research Centre Europe (France)

N. Mamede e J. Baptista
L2f/INESC-ID Lisboa (Portugal)