Compreensão de linguagem natural – voltando à carga

Diana Santos

Agradecimentos

Ao António Teixeira por me ter perguntado:
– O que andas a fazer?

Ao Eckhard Bick pelo uso do PALAVRAS

Ao resto dos elementos da Linguateca ☺

Aos membros do RSC/VD pelo uso do **titan**
(High Performance Computing cluster da Univ. Oslo)

Aos variados órgãos de financiamento da Linguateca e meus

POS_Conhecimento
Programa Operacional Sociedade do Conhecimento

União Europeia
FEDER

UMIC
Agência para a Sociedade do Conhecimento

FCCN
Fundação para a Computação Científica Nacional

MCTES MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

O MUNDO E A LÍNGUA

Conhecimento do mundo

- Relação íntima entre conhecer e pensar e falar
- Sem uma língua não é possível pensar (= desenvolver categorias, generalizar, identificar, perceber)
- Antes de pensar não é possível falar, por isso a comunicação é um sub-objetivo da língua
- Sem mundo não há conhecimento

A língua tem forma e conteúdo

- O conteúdo não se esgota em informação
- De facto, o principal conteúdo é emoção
- Línguas diferentes proporcionam um conhecimento diferente do mundo (posição whorfiana)

Relação com a IA

- Representar o conhecimento "independentemente da língua" é um disparate
- Traduzir para uma linguagem "melhor" que a LN é uma quimera
- Os "problemas" da LN são a sua vantagem competitiva

O que se pode fazer

- Escolher formas de simplificação e de generalização para objetivos concretos
- Processá-las / "raciocinar" com elas de forma a criar novo conhecimento (enriquecer a nossa visão do mundo) que irá depois refletir-se na LN ("quarks")

É tão importante o que se diz como o que não se diz

- Carta de recomendação:
  - *Chega sempre a horas, não bebe*
- Não obtive resposta
- *vermelho*
  - Quando é que se diz que algo é vermelho? Quais as coisas mais vermelhas?
- *medicina*
  - onde é que esta palavra aparece mais?

Dados dos corpos e a língua

- Diferença entre o que se diz e o que se sabe
- Diferença entre o que é culturalmente esperável e o chocante
- Diferença entre criação e repetição
- Humor, ironia, e sarcasmo

Estruturalismo

Para perceber o que se diz

- é preciso saber as alternativas, o sistema
- é preciso distinguir entre escolha e rotina
- é preciso atender também às conotações
- é preciso estar consciente do contexto

O método de Veale

- Obter automaticamente o "conhecimento consensual", que as pessoas têm sem nunca o proferir explicitamente (*common sense*)
- Através das perguntas no Google
  - *Porque é que os gatos fogem?*
  - *Porque é que as frigideiras de Teflon*
- Através das comparações diretas

A minha caracterização da linguagem natural

PROPOR 2006

1. Natureza metafórica
2. Dependência do contexto
3. Referência a conhecimento implícito
4. Vagueza
5. Caráter dinâmico (evolução e possibilidade de ser aprendida)

Relação entre forma e conteúdo

- Toda a língua tem uma forma e um conteúdo
- Mas se a forma é "objetiva", o conteúdo escapa-se-nos entre os dedos, ou... é negociado através da interação (com o resto dos falantes e com o contexto)
- A língua escrita é mais fácil de entender/preservar porque tem menos contexto não linguístico

Notas de alta

- Toda a gente entende? Não
- Para que servem?
  - Para ajudar os médicos
  - Para controlar os médicos
  - Para ajudar o doente
  - Para fazer investigação (prospeção)
  - Para fazer história da vida quotidiana

Método de Schank

- Um guião:
- Entra doente – é interrogado – fazem-lhe exames – dão-lhe tratamento – vêem se melhora – mudam tratamento – sai doente

A questão das ontologias e da web "semântica"

- Tudo se pode traduzir em triplos

Entidade1 relação entidade2

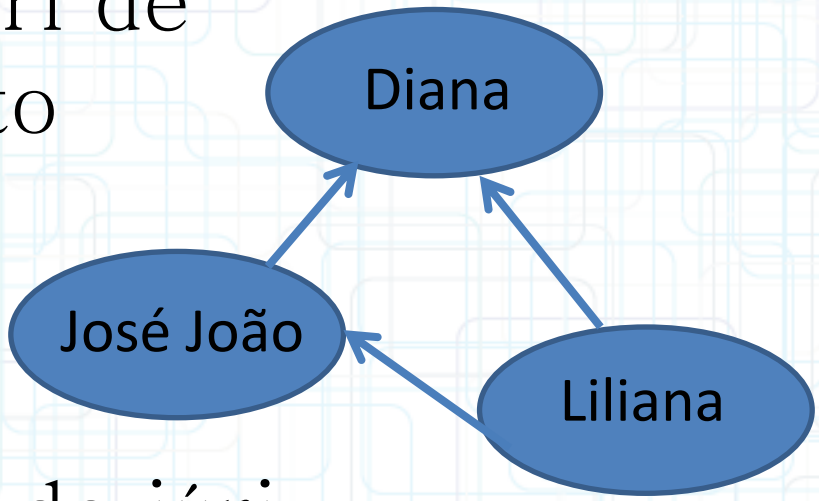
- Mas: nem todas as relações têm um nome (na língua)
- A maior parte das relações dependem de outros conceitos/relações
  - X irmão-de Y
  - X colega-de Y

As relações não são simétricas

- X professor-de Y (grupo de alunos, turma)
- X padrinho de Y (grupo de afilhados)
- X marido de Y (num harem)
- X ama-de-leite Y (irmãos **colaços**)
- Co- co-orientadores, co-ordenação, co-operação, co-piloto, co-adjuvante

As relações são dinâmicas

- Teve no júri de doutoramento



- Foi colega de júri
- ser aluno de / fui aluno de
- E têm uma importância diferente para cada interveniente

Mas há casos óbvios ☹

- Lisboa capital de Portugal
- Relação? (de que tipo?)
- Não, é uma propriedade (única) de uma instância relativa a outra, um papel que é atribuído (num país)
- Propriedade geográfica?
- Não, é uma propriedade histórica, pelo menos como intensão.

Local ou patologia?

- Insuficiência renal
  - Nos rins ou DOS rins?
- Fractura cervical
- Alergia crónica
- Problema digestivo
  - Em que parte do aparelho digestivo?
- Expetoração seca

Uma ontologia...

- É uma forma de organizar o conhecimento, para um objetivo concreto
- Tem de ser uma simplificação da linguagem natural (a nossa linguagem de lidar com o conhecimento) que permita estruturar o que já sabemos
- Mas é preciso cuidado para não esquecer o que já sabemos

À laia de conclusão

- Não esquecer que a nossa linguagem natural é o português
- Não acreditar em conhecimento independente da língua: quem define os nomes das doenças? A doença do pezinho já tem nome em inglês?
- Não esquecer que domínio da língua é poder – e a terminologia não é neutra

E EU?

Do que vou falar

Do que se aprende ao ensinar uma língua

Anotação semântica

O Págico (português mágico)

Problemas “reais”

Avaliação

Ensino do português a nível universitário

- Ensinar gramática
- Pressupondo
  - que eles sabem falar
  - que eles sabem gramática
- Numa altura em que a exposição ao português não é difícil
  - Na rede, e "viagens de língua"

Juntando as duas questões

1. Descrição da gramática da língua para o ensino
2. Descrição da gramática da língua para o PLN

= Uso dos corpos anotados da Linguateca para fazer uma descrição empírica do português

Quantidades no grupo AC/DC

- 21 corpos, 295 milhões de unidades
- Variante: 224.254.595 PT e 65.671.800 BR
- Género: 252 milhões jornalísticos, 17 milhões literário, 4,5 milhões técnico
- Por artigo (no caso jornalístico de artigos completos): 245.490 artigos
- Frases: 12.639.914

Dados de Julho de 2011

"Banho de língua"

O primeiro sistema desenvolvido foi o *Ensinador*: Simões & Santos (2011)

- Material autêntico
- Exercícios à vontade do freguês/professor
- Escolha de bons exemplos e de géneros e variantes apropriados
- Resultados imediatos
- Principal vantagem: ilustra diferentes razões de escolha!

Págico: avaliação conjunta de recolha de informação na wikipédia portuguesa

Sistemas automáticos de obtenção de informação não-trivial em português, sobre cultura lusófona

Avaliação indireta da wikipédia

Chamada de atenção sobre a cultura em português

Insistência em sistemas úteis para um público maior

Problemas da wikipédia

Não é um sítio onde TODA a informação se encontra

- Há mais na área da ficção científica
- Há mais na área da informática
- Há mais na cultura anglo-saxónica / anglo-americana

Um exemplo

Marcel Proust

pt.wikipedia.org/wiki/Marcel_Proust [21Jun2011]

Ela era culta e bem informada; suas cartas demonstram um senso bem desenvolvido de humor e seu domínio do inglês foi suficiente para lhe fornecer a assistência necessária para as tentativas posteriores de seu filho de traduzir John Ruskin.

Um exemplo

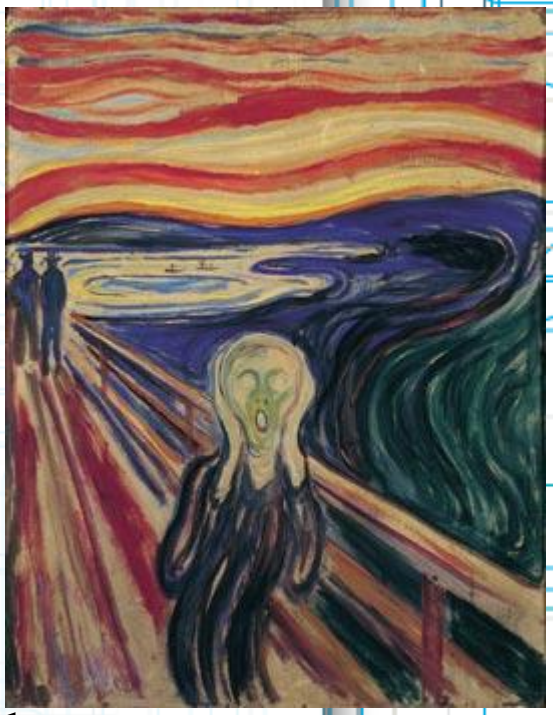
Marcel Proust

en.wikipedia.org/wiki/Marcel_Proust [21Jun2011]

She was literate and well-read; her letters demonstrate a well-developed sense of humour, and her command of English was sufficient for her to provide the necessary assistance to her son's later attempts to translate John Ruskin.

Problemas reais

- A linguagem natural é a forma de exprimirmos o nosso conhecimento
- É a forma como comunicamos e nos mal-entendemos
- É o instrumento com que trabalhamos a maior parte do tempo



“Colaboração” com o RCAAP

- RCAAP, http://rcaap.pt/
- Intervenientes: KEEP, SDUM, FCCN
- Será que o PLP pode ajudar no desenvolvimento, ajuda à entrada de dados, validação, ou sugerir novas funcionalidades?
- Conhecimento semi-estruturado em português: nomes de autores, de instituições, “locais”, assuntos

Será?

Para responder a esta questão, temos de ter acesso a:

- O que lá está: Universo dos metadados (recebemos irregularmente por cortesia da KEEP)
- O que se procura: Universo da procura (diários da interação dos utilizadores) – só é possível no meta-repositório, e mesmo assim com limites (levantamos regularmente)

Calendário

- Ideia aceite e iniciada em Outubro de 2009
- Primeiras ideias publicadas em 2010 (Santos & Ribeiro, 2010)

  Universo dos nomes de autores

  Ambiguação dos nomes de autores
- Metodologia e dados atuais em Junho de 2011

Estudo de sessões no RCAAP

- Dificuldade de definir um utilizador
  - Mesmo IP e agente é um critério fraco demais → muitos utilizadores da mesma instituição são amalgamados
  - Manutenção do IP para sessões diferentes é outra heurística desajustada à realidade atual → muitos utilizadores que acedem ao RCAAP de pontos diferentes não são identificados como o mesmo

Informação sobre os utilizadores

- São investigadores à procura de artigos, ou são as pessoas que acabaram de depositar os seus dados e querem confirmar que estão bem?
- Para quem é o RCAAP? Para quem foi pensado o RCAAP? Quem utiliza o RCAAP, de facto? Os bibliotecários? Ou os burocratas?

http://www.rcaap.pt/help.jsp

O RCAAP é portal agregador (meta-repositório) que reúne a descrição (metadados) dos **documentos depositados nos vários repositórios institucionais em Portugal**. O Portal RCAAP recolhe o texto integral para melhorar o resultado das pesquisas mas não guarda qualquer documento.

- Para além de poder **pesquisar a produção científica portuguesa**, pode optar por pesquisar também a produção científica brasileira que neste momento é composta por vários repositórios e revistas agregados no projecto OASIS.

Tipos de procuras

- Por autor
- Por título
- Por assunto
- Sem precisar (=procura simples)

Resultados das procuras

- Distribuição dos resultados obtidos
- Bom ou mau?
- Quantas vezes se refinou?
- Se não se obteve nenhum resultado
- Se se obteve resultados
  - aumentou
  - diminuiu

Descrição do material

- Quantos autores são ambíguos (= passíveis de terem várias interpretações)?
- Quantas vezes os autores "ambíguos" têm de ser re-procurados/refinados?
- Qual é a melhor maneira de os distinguir para um utilizador? (assunto(s), data de publicações, inst., co-autores, ...)?

Como se avalia se vale a pena fazer esse sistema?

- A posteriori, se for usado...
- E se o sistema não for possível? Ou seja, se não houver suficiente informação no RCAAP para separar corretamente autores diferentes?
- E se o sistema tiver 10% de erros, no sentido de que considera como autores diferentes as mesmas pessoas?

Precisão ou abrangência?

Procura guiada, ajuda ao utilizador:

- Existem X autores com esse nome que
- publicaram sobre os seguintes assuntos: escolha os que (não) lhe interessam
- publicaram pelas seguintes instituições...

Citações do Eckhard ou do PALAVRAS

E se quisermos apreciar ou medir o impacto de uma pessoa num conjunto de publicações?

Imaginando que temos acesso, já não só aos metadados sobre as publicações, mas ao texto dessas mesmas publicações (ou às citações):

Poderíamos medir, nas áreas relacionadas, quantas citações existem. Isso é o que a bibliometria faz ou tem vindo a fazer há muitos anos. Mas uma lista é suficiente?

Como medir a importância?

- *O nosso sistema é semelhante a Bick (2000)*
- *Outros sistemas, tal como Bick (2000)*
- *Ao contrário de Bick (2000)*

- *Usámos Bick (2000)*
- *O nosso sistema é baseado em Bick (2000)*
- *A nossa pesquisa usa o AC/DC*

Em conclusão

- Estamos, no PLP, a quilómetros de distância de um sistema útil para fazer coisas com sentido
- ... mesmo que estejamos apenas no subconjunto restritíssimo de nomes de autores
- Ou porque estamos nesse conjunto restrito?