Objectivos do III simpósio doutoral da Linguateca, comentários e pontes

Diana Santos
Linguateca
www.linguateca.pt

História dos objectivos

- Acompanhamento do trabalho
  - encontros presenciais
  - necessidade de relatar o progresso e as ideias (tb as falhadas)
- Crítica e aperfeiçoamento das ideias
  - através da presença da orientadora
  - através de estímulo conjunto à crítica
- Desenho conjunto do futuro
  - Através do estabelecimento de pontes
  - Fertilização cruzada
- Disseminação da Linguateca aos próximos

Método

- Apresentação do trabalho feito
- Teste e apresentação do estado da arte numa dada área
- Apresentações de interesse geral
  - Sobre técnicas ou tecnologias
  - Sobre assuntos relevantes
  - Sobre avaliação

As minhas contribuições desta vez

- Alguns comentários sobre estados da arte e referências
- O que é semântica?
- O papel das pessoas em PLN
- Esclarecimento de algumas (presumíveis) confusões
  - Conhecimento cristalizado vs. em contexto
  - Informação vs. Conhecimento
  - Web semântica e Web 2.0
- Tentativa de maior colaboração entre os participantes
- Tentativa de evitar alguns "erros"
  - Servir para tudo, limites fixos, "funcionar" antes de fixar os critérios

Referências e estado da arte

- Idealmente
- Visão original sobre o assunto
- Enviesamento da apresentação dos autores

- Pelo menos
- Evidenciar conhecimento dos maiores e das principais estados da arte já feitos (referindo-os)
- Salientar os conceitos mais importantes e que vão ser usados no resto do texto
- Comparar terminologia esclarecendo-a

conferir mails de orientação

O que é semântica?

- Relação com o mundo exterior

E

- Relação com as nossas mentes

- O que significa compreender?
  - Modificar as nossas mentes
  - Conhecer melhor o mundo exterior
  - Gerar nova informação entre as nossas mentes e o mundo
- O que significa aprender?
  - Criar conceitos na mente relacionados com fenómenos no mundo exterior
  - Relacionar esses conceitos

A semântica e a sintaxe são duas faces do mesmo

- Estrutura de conhecimento que permite a aprendizagem
- Sintaxe: dá regras, pistas para compreender, automatiza o processo de ouvir e de falar
- Semântica: apoia-se na sintaxe para não ter de dizer tudo de cada vez

- Pragmática: o cesto do lixo de alguns investigadores: semântica sem sintaxe (entoação, contexto…)
  - Disparate definir sintaxe sem entrar em conta com fenómenos da língua falada também…
  - Disparate separar arbitrariamente a linguística em campos diferentes

O que é semântica computacional?

- Mundo exterior pode ser modelado dentro do computador
  - Bases de dados, sistemas periciais, robôs
- As nossas mentes podem ser modeladas dentro do computador
  - Conceitos, regras, crenças, inconsistências, ignorância, aprendizagem

- Duas visões diferentes?
  - Substituição do humano: Web semântica
  - Ajuda ao humano: Sistemas de apoio à decisão

Diferenças entre SW e Web 2.0

- Formalização para computadores dispensarem humanos
  - estrutura rígida/artificial
  - pessoas libertadas do trabalho chato
- "Informalização" para mais humanos participarem em computadores: juntar o humano ao computador numa sociedade maior
  - estrutura espontânea/emergente
  - pessoas convidadas a trabalhar por gosto

- É possível "dar a César o que é de César" e ter o melhor de dois mundos?

Serviços: terminologia

- Serviços na Web: para pessoas
  - COMPARA
  - Ciberdúvidas
  - Esfinge
  - AsjJeeves
- Serviços Web: para programas
  - Google API
  - Amazon API
- Serviços Web semânticos (ou serviços da Web semântica)
  - ?? em que está envolvido raciocínio
- Agentes?? programas que executam tarefas semânticas, podem invocar SWS...

Dica 1: Deitar fora as constantes

- O que significa fazer programas que dão
  - um número constante de significados para uma palavra?
  - um número constante de traduções para uma palavra?
- Alternativa:
  - dependendo da palavra e das condições em que ela se encontra, fazer um número variável de significados, traduções, relativo a parâmetros relevantes (por exemplo 90% da frequência, ou algo mais complicado)

Dica 2: Perceber a diferença entre tipo e realização (token)

- Tipo: sem contexto
- Token: em contexto

- Dicionários (monolingues, de tradução, etc.): ligam tipos
- Instâncias (corpora, corpora bilingues): ligam tokens

- Probabilidades de ocorrência: referem-se a tipos (abstracções)
- Probabilidades de instâncias: não faz sentido!

Perceber a diferença entre um corpus e um dicionário

- corpus: real
- dicionário: abstracção

- A: classificação sintáctica num corpus: a classificação daquela instância
- B: classificação sintáctica num dicionário: os vários potenciais de uma instância, sem que o contexto seja demasiado relevante

- Ou seja A pode ser um dos vários potenciais (elementos de B), ou outra coisa!

Digressão HAREM: Diferença entre um almanaque e REM

- almanaque: tipos
- REM: análise de texto
- almanaque mais bem desenhado: tipos com vários potenciais

- Não é necessário que os tipos do almanaque sejam os mesmos do REM
- almanaque: *França* – país; *Diário de Notícias* – jornal
- nível intermédio de semântica: um país é, simultaneamente e em abstracto, um conjunto de coisas (indissociáveis)
- Em REM, na prática, pode apenas um desses sentidos ser activado

Nível intermédio de semântica (Cruse 2004)

- Estudo da polissemia: sentidos relacionados (sistematicamente)
  - relações lineares: autohiponímia, autohiperonímia, automeronímia, autoholonímia
  - *proibida entrada de cães; a história do homem; pinta/tira a porta; ela arranhou/perdeu um braço*
  - relações não lineares: metáfora (semelhança), metonímia (associação)
- Entre a polissemia e a monosemia...
  - facetas (partes discretas mas não antagónicas) de um único conceito
  - *livro, banco*
  - perspectivas: qualia: papéis constitutivo (partes), formal (classe), télico (função), agentivo (origens)
  - *cavalo*

Semântica em contexto (Cruse 2004)

- **Selecção**: um dos que está listado nos tipos
- **Coacção**: procurar extensões (metáfora, metonímia) que permitam compreender
- **Modulação contextual** (dentro dos limites do sentido)
  - enriquecimento: torná-lo mais específico. *A casa foi roubada. Só **levaram** o X*
  - empobrecimento: *as crianças formaram um **círculo** à volta da professora*

Voltando ao REM

- Vários níveis que podem ser desejados
- e que respondem a perguntas diferentes

- MUC: quantas vezes é que a palavra *France* foi usada como país?
  - Perde-se a informação de que tipos de significado (lugar, organização, pessoas, ideia) são atribuídos a um país
- HAREM: quantas vezes é que a palavra *França* foi usada para indicar um lugar?
  - Perde-se a informação de a que tipo mais geral França pertence (país, freguesia, sala)
- São possivelmente complementares (desde que não se faça o erro de postular, e acreditar que país, freguesia = lugar)

Motivação

- MUC: Mais fácil obter dados objectivos, mais simples
- HAREM: Mais próximo da compreensão de um texto, mais próximo das necessidades dos utilizadores/aplicações

- Interessante: comparar as duas abordagens, por exemplo aplicando/revendo a categorização tipo MUC à colecção dourada do HAREM (e vice-versa, se alguém quisesse trabalhar para o inglês)

- Discussão: interacção com a análise sintáctica/semântica
  - Eckhard: Sabendo o tipo MUC e dado o contexto sint-sem é possível derivar o tipo HAREM (??)

Mais ideias

- Sabendo a distribuição das classificações do tipo HAREM é possível obter os tipos MUC (hipótese maluca)?
- Seria um estudo interessante... (semelhante ao reagrupamento de forma a obter classes morfológicas e/ou sintácticas)

- Ideia geral:
- É preciso estar sempre a saltar entre tokens e tipos, usando a informação de uns para chegar aos outros, para voltar aos primeiros, para melhorar a extracção dos segundos ...

Pontes: mais uma vez...

- Extracção de informação -> extracção de conhecimento

- classes e instâncias
- gentílicos, objectos, ferramentas, frutas, doenças, ...

- tipos e tokens – diferente de classe-instância (sempre a nível de tipos)

- a primeira classificação é o nome: criação de um tipo
- *O Fiel é o meu cão*
- O tipo *Fiel* passa a designar sempre o mesmo objecto/animal no meu contexto

Tipos e menções/realizações

- Mas cada tipo básico pode ser atribuído a um conjunto **infinito** de classes
- cão (Fiel), animal(Fiel), amigo(Diana, Fiel), mamífero(Fiel)...
- 2 casos: tipos com o mesmo nome, tipos com outro nome

- Paris – tipo, Paris – elemento da classe cidades, Paris – pessoa mit.
- cidade(Paris) – um objecto no mundo
- *Paris*: milhares de invocações desse mesmo objecto + milhares de invocações de outros objectos. Em cada contexto é um token diferente

Dica 3: Problema concreto e solução em vaivém

- Medir quantas EMs (nomes próprios no sentido do HAREM) têm a ver com LUGAR na Web portuguesa
  - primeira estimativa
    - medição do erro (erros e faltas)
    - análises dos erros
  - melhoria das ferramentas de medição (SIEMES, REPENTINO, SEI?)
  - nova estimativa
    - medição do erro ... etc

Problema concreto e solução em vaivém

- Medir a correcção do alinhamento das palavras no COMPARA
  - primeira estimativa
    - medição do erro (erros e faltas)
    - análises dos erros
  - melhoria das ferramentas de medição (NATools, Apalinha)
  - nova estimativa
    - medição do erro ... etc

A relação entre as pessoas e a máquina ☺

- Todos os programas / sistemas feitos em PLN são ultimamente para satisfazer pessoas (porque senão não era preciso PLN)
- Os avaliadores derradeiros de sistemas de PLN são sempre pessoas
- Errar é humano (e computacional)
  - É importante entrar em conta com esse erro
- Julgar é humano (e computacional?) e as pessoas divergem
  - É importante saber entrar em conta com essa divergência
- As pessoas mudam
  - É importante saber adaptar programas à disposição das pessoas que com eles lidam

Irrar é humano

- Os programas têm de ser robustos e
  - contar com erros ortográficos, sintácticos, semânticos, lógicos, de tradução etc.
  - ajudar a detectar e corrigir os erros
  - deixar que os humanos insistam nos erros
- Os programas têm de ser robustos de forma a não serem confundidos pelos erros
  - ao generalizar
  - ao inventariar
  - ao "raciocinar"/traduzir
- Os programas não podem ser cegamente avaliados por comparação com o desempenho humano

Julgar é humano

- Atitudes, opiniões, estados de espírito, "feelings"
- Não adianta o computador ter razão se as pessoas não lha dão...
- Interessa saber comparar opiniões (de pessoas)
  - concordância inter-anotadores
  - concordância por classe
- Nem sempre interessa haver concordância inter-anotadores!
  - sistemas personalizados deveriam discordar tanto quanto as pessoas a quem se personalizaram...
- O mito da objectividade é apenas o mito do consenso:
  - quando há muitas pessoas que concordam, é "objectivo"
  - quando há muitas pessoas que discordam, é "subjectivo"

Julgar é humano (2)

- Atitudes, opiniões, estados de espírito, "feelings"
- Há graus de confiança

- classificação fora do contexto
- *pena*
- substantivo
- *Que pena!*

- "substantivo em contexto"
- *a pena de morte, a pena foi aliviada, sem pena nenhuma, pena do chapéu, Um professor é um responsável que pena!*

Informação vs. conhecimento

- Informação: estática
- Conhecimento: informação que pode ser aplicada
  - geralmente pressupõe a capacidade de raciocinar sobre essa mesma informação

Lista de propriedades da LN que a tornam diferente das outras línguas

1. Natureza metafórica
2. Dependência do contexto
3. Referência a conhecimento implícito
4. Vagueza
5. Carácter dinâmico (evolução e capacidade de aprendizagem)
6. Interacção (diálogo)

Três tipos de vagueza

1. monodimensional, limites difusos
- quantos cabelos se pode ter e ainda ser *careca*?

2. um conjunto de propriedades em eixos diferentes
- cidadão sueco: pais suecos E nascido na Suécia
- substantivo: género inerente; sem grau; contável, massivo ou abstracto

3. conjunto de propriedades relacionadas
- classe das aquisições (o estado e a sua incepção)
- oposição privativa
  - gosto: bom gosto, mau gosto
  - sorte: boa sorte, má sorte

Ligação entre as várias propriedades

- Vagueza e dependência de contexto explicam mudanças de significado e daí, evolução da língua
- Diálogo explica como a língua pode ser aprendida, e como as pessoas aprendem negociando o sentido
- A língua baseia-se em conhecimento implícito, e desenvolve-se no sentido das necessidades de comunicação de uma comunidade, o que explica divergência e convergência das línguas
- Metáforas partilhadas permitem mudanças de sentido sem traumas

Várias línguas (linguagens naturais)

- As línguas são genuinamente diferentes
  - reflectem uma visão diferente do mundo
  - contêm uma "cola" diferente (sintaxe, discurso)
  - entram em conta com diferente informação implícita
- Uma teoria do funcionamento da linguagem natural e de como evolui devia explicar este facto indiscutível
  - impossível aprender outra língua materna por um adulto
  - muito difícil fazer tradução automática
- Tratar da sua língua tem um impacto político apreciável

Línguas diferentes: causas

- contextos diferentes; falantes diferentes
- informação implícita diferente; constante uso creativo

Admira que as línguas divirjam?

- *it would surely be surprising, and a very strong empirical claim, that different languages using different means to express 'meanings' always arrived at exactly the same end*

Keenan, Edward. "Some Logical Problems in Translation", in Guenthner & Guenthner-Reutter (eds.), *Meaning and Translation: Philosophical and Linguistic Approaches*, Duckworth, 1978, pp.157-89.

Línguas diferentes: resultados / factos

- gramática diferente
- itens lexicais diferentes
- convenções diferentes sobre implícito/explícito
- regras diferentes para obrigatoriedade e opcionalidade
- diferentes estratégias discursivas
- coisas diferentes que fazemos com as palavras
- diferentes métáforas convencionalizadas
- diferentes realidades descritas/facilmente invocadas
- diferentes provérbios/cultura