Avaliação: fundamentos e sua aplicação ao PLN

Diana Santos
Primeira Escola de Verão da Linguateca

Objectivo do módulo

- Questões principais; métodos e resultados.
- O paradigma da avaliação conjunta.
- Casos concretos, relacionados com os módulos da escola.
- Avaliação de sumarização em português, de recolha de informação, de análise sintáctica e morfológica.

Avaliar para quê?

- Essência do método científico
- Essência da engenharia
- Essência da condição humana: valores
- Na prática: a avaliação é pouco reconhecida e não é vista como suficientemente meritória pela sociedade e/ou pelos órgãos de poder

Avaliar em PLN

Dado que o processamento de linguagem natural

- se baseia em modelos de como a língua funciona
- desenvolve aplicações que lidam com ela
- para funcionarem com seres humanos

há 3 focos fundamentais de avaliação em PLN

Vários tipos de avaliação

- Avaliar uma hipótese na prática (implementando aquilo que parece possível)
- Avaliar se um sistema ou serviço faz aquilo para o qual foi destinado
- Avaliar se os utilizadores o conseguem usar
- Avaliar se o desenvolvimento e sua introdução na vida real foi positivo

Os ingredientes de uma avaliação (técnica)

- Uma descrição ideal, suficientemente concreta para ser testável
- Um conjunto suficiente de casos para os quais se sabe a resposta, ou de condições diferentes em que o sistema deve dar a mesma resposta, ou de precisão de resposta...
- Uma forma de comparar a resposta do sistema com a resposta pretendida

Exemplos

- Uma calculadora electrónica, um parser
  - experimentar algumas contas
  - experimentar a análise de algumas frases
- Um rádio, um parafraseador
  - experimentar em várias altitudes
  - experimentar diferentes maneiras de dizer o mesmo
- Lâmpadas, um sistema de RI ou de RAP
  - 9 em 10 sejam adequados

Avaliação centrada no usuário

- 9 em 10 utilizadores contentes?
- Fazer uma tarefa com o sistema
  - antes e depois
- Aprender
- Tempo de começar a trabalhar
- Empatia

- Muito mais difícil e custoso

Medidas

O que é uma medida?

- objectiva
- repetível
- independente do contexto
- ordenável
- compreensível
- intuitivamente relacionada com qualidade

RMG ("baseline") e tecto ("ceiling")

- Se um sistema não faz mais do que um RMG (rendimento mínimo garantido), é inútil
  - sistema que associa a cada palavra a sua categoria gramatical: sempre N
- Há um tecto a partir do qual não é possível medir, porque não há consenso
  - se 10% dos casos não conseguem ser decididos por seres humanos, o tecto é 90%

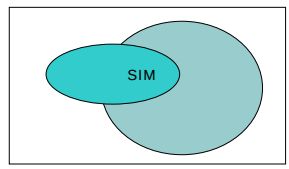
Modelos de avaliação em PLN

- avaliar modelos (Baayen 1993)
- avaliar uma hipótese de desempenho (Koeling et al 2005)
- avaliar uma medida ou representação (Sampson & Babarczy 2003)
- avaliar a execução de uma tarefa
- Avaliação conjunta (evaluation contest):
  - MUC, HAREM, TREC e QA/CLEF, DUC

Algumas medidas e conceitos

- Precisão = Sim/(todo o resultado do sistema)
- Abrangência = Sim/(tudo o que há)

- Relevância
- Correcção
- Utilidade
- Semelhança
- Novidade, ...



Formas de comparar

- Com algo certo que se arranja à mão antes (recursos dourados)
- Com algo certo que se corrige depois (monte, "pool", correcção)
- Com algo certo que se estraga (mutilação)
- Integrado numa coisa maior
  - uso pelos utilizadores
  - tarefa mais abrangente

Quase todos os conceitos são

- complexos e vagos
- é mais fácil defini-los do que verificá-los

- ambiguidade
- vagueza
- homonímia
- polisemia

Vagueza e ambiguidade

- V: A mesma unidade significa ao mesmo tempo várias coisas relacionadas entre si.
- Diferenças essenciais em relação a A:
  - relação entre as traduções
  - a vagueza é **sistemática**, ambiguidade é acidental
  - V é relevante para a comunicação, A não
  - na tradução entre duas línguas, perde-se informação se não se preserva V

Como avaliar modelos

- Baayen, R. H. "Statistical models for word frequency distributions: A linguistic evaluation", *Computers and the Humanities* **26** (1993), pp. 347–363.

Avaliação de modelos estatísticos em função de adequação linguística (Baayen 93)

- Produtividade morfológica: prontidão estatística para formar espontaneamente novas palavras
- Qual o tamanho do vocabulário teórico?
- "Word frequency distributions": distribuição da frequência das palavras
- Para, a partir de contagens finitas, poder generalizar

Como avaliar as distribuições?

- Ajuste aos dados empíricos
- Fundamentação linguística
- Valores em questão (resumir frequências):
  - distribuição ordem-frequência $freq_i = f(rank_i)$
  - distribuição de frequência agrupada: uma classe para todos os elementos com a mesma frequência
- V(N):vocabulário observado numa amostra N
- $n_r(N)$: número de tipos com frequência r numa amostra de tamanho N

Que distribuições?

- Lei de Zipf generalizada
- Lei lognormal
- Lei inversa de Gauss-Poisson generalizada

- $X^2_{N,K}=(x-\mu)^T\sigma_{ij}^{-1}(x-\mu)$
- x: (V(N), $n_1(N)$, $n_2(N)$, ... ,$n_k(N)$)
- μ: E[V(N)], E[$n_1(N)$], E[$n_2(N)$], ... , E[$n_k(N)$])

Resultados

- Cobuild corpus (15,7 milhões de palavras)
- 1 obra de Puschkin
- Substantivos monomorfémicos em holandês
- 4 sufixos derivacionais em holandês *(-je, -ing, -er, -heid*)

- lognormal modela texto literário mas não corpora; modela sufixos produtivos mas não idiomáticos; tem fundamentação linguística

Resultados (cont.)

- Lei inversa de Gauss-Poisson generalizada só modela bem os nomes monomorfémicos e não tem fundamentação linguística
- Lei de Zipf generalizada tem mais fundamentação linguística, modela melhor corpora... Mas:
- nenhuma das fundamentações entra em conta com diferenças entre palavras morfologicamente simples e complexas!

Resultados sobre o tamanho do vocabulário

- Escolhendo a distribuição que mais se ajusta a cada caso, e arranjando uma estimativa externa:
- Cobuild: número de lemas nos dicionários
- Puschkin: vocabulário da obra do autor
- Morfologia: contando na base CELEX (dicionário)

Avaliar uma hipótese de desempenho

- Koeling, Rob, Diana McCarthy & John Carroll. "Domain-Specific Sense Distributions and Predominant Sense Acquisition". In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. HLT/EMNLP 2005*, pp. 419-426.

Comparar um método

- Adquirir informação sobre qual o sentido predominante de uma palavra
- para construir corpora específicos anotados semanticamente
- para treinar sistemas de desambiguação de sentidos (WSD)

- RMG (baseline) no Senseval 3: escolher sempre o sentido mais frequente no SemCor

Porquê

- um sentido por discurso (Gale et al 1992)
- domínio do texto é relevante (Magnini et al 02)
- tinham um método para obter esse sentido de texto não anotado:
  - extraem um tesauro de cada texto, analisando-o gramaticalmente – inventário de sentidos: WordNet; semelhança: Lin (1998); 50 vizinhos por palavra; cada sentido recebe uma pontuação com base no método **jcn** que usa frequências extraídas do BNC

Preparativos para testar a hipótese: escolha de 3 tipos de texto

- BNC escrito (90 milhões de palavras (Mp), 3209 documentos)
- FINANCE: 117.734 textos sobre finanças extraídos do corpus da Reuters (32,5 Mp)
- SPORTS: 35.317 textos sobre desporto (9,1Mp)
- Criaram os tesauros para estes 3 corpora

Preparativos para testar a hipótese: escolha das palavras

- Usando os códigos de assunto na WordNet, extrairam todas as palavras (nomes)
  - que tinham sentidos tanto em finanças como em desporto (38)
  - com frequência maior do que 1000 no BNC
  - com no máximo 12 sentidos diferentes
  - pelo menos 75 exemplos em cada corpus
  - retiraram alguns com sentidos muito esquisitos
- 17 palavras, mais...

Preparativos para testar a hipótese: escolha de mais palavras

- Mais 3 grupos de palavras
  - particularmente salientes em SPORTS (8)
  - particularmente salientes em FINANCE (8)
  - igualmente salientes nos dois (7)
- Saliência da palavra *w* no domínio *d*: $S(w,d)=(N_{wd}/N_d)/(N_w/N)$
- A escolha foi feita aleatoriamente entre as 50 mais salientes de cada grupo
- Em média, as 40 palavras têm 6,6 sentidos

Preparativos para testar a hipótese: anotação manual dos sentidos

- 10 anotadores de língua inglesa
- Aparecem os sentidos (aleatoriamente): definição, e exemplos
- só vêm uma frase. As frases são obtidas aleatoriamente dos corpora
  - Se não conseguem decidir, devem escolher "Unclear"
  - se o sentido é outro, devem escolher "Falta na lista"
- 33.225 anotações, cada frase foi anotada por 3 ou mais, 65% de concordância entre anotadores

A distribuição dos sentidos

- Entropia da distribuição de um sentido é uma fracção da entropia total
- i varia entre todos os sentidos
- $\text{Entropia}=(-\Sigma_i\, p(i)\log_2 p(i))/\log_2(\#\text{sentidos})$
- Frequência relativa do sentido dominante

bill	Entropia	rel(fs)	sentido
BNC	0,503	42,61	1
FINANCE	0,284	77,0	1
SPORTS	0,478	45,2	2

Conclusões

- Avaliar o método de atribuir o sentido certo
  - Ir buscar o sentido mais frequente ao SemCor
  - Método executado sobre o texto não anotado
- RMG (baseline): Σ_i 1/(#sentidos (i))
- Comparam
  - treinar no mesmo domínio ou noutro
  - treinar no SemCor
- Quando há "diferenças substanciais" entre domínios, treinar nos ditos é melhor, visto que estes corpora exibem maior entropia do que o BNC ou o SemCor

Avaliar uma medida ou representação

Questão inicial: Black et al PARSEVAL, GEIG

- Únicos constituintes consensuais (9 gram.):
- *The famed Yankee Clipper, now retired, has been assisting (as (a batting coach)).*
- Preparação
  - Cada sintagma deve ser incluído entre ()
  - Várias regras precisas para transformar os ()
- Não ligar a constituintes unários

Como fazer a comparação

- Comparação com o Penn Treebank (cont.)
  - Número de constituintes incompatíveis (CI)
  - Precisão em termos de constituintes
  - Abrangência em termos de constituintes
- Combinação
  - Distribuição do número de CI
  - Precisão média
  - Abrangência média

Comparar o desempenho de medidas

- Sampson, Geoffrey & Anna Babarczy. "A test of the leaf–ancestor metric for parse accuracy", *Journal of Natural Language Engineering* **9**, 2003, pp. 365–80.

"leaf-ancestor": diferença entre as linhagens (caminhos entre as folhas, palavras, e a raiz da árvore)

Ilustração da "nova" medida

[S [N1 *two* [N1 *tax revision*] *bills*] *were passed*]

[S [NP *two tax revision bills*] *were passed*]

- *two* N1 [S: NP [S
- *tax* [N1 N1 S: NP S
- *revision* N1] N1 S : NP S
- *bills* N1] S : NP] S
- *were* S : S
- *passed* S] : S]

Ilustração da "nova" medida

- Semelhança entre duas linhagens recorre à distância de edição (Levenshtein) Lv (1 por cada operação Ins, Apag, Subst)
- 1-Lv(cand,dour)/(comp(cand)+comp(dour))
- Subst=f com valores no intervalo {0,2}
  - Se a categoria é relacionada (partilha a mesma primeira letra), f=0,5, senão f=2
- Numa frase a semelhança é dada pela média

Comparação com outras métricas

- Aplicaram a 500 frases do SUSANNE (aleat)
- A ordenação é diferente
- Não há correlação entre GEIG com nomes e sem nomes
- Exemplos concretos de diferenças extremas em que os novos condizem com a intuição
- Possibilidade de apontar onde se encontra o problema (vendo as classificações por palavras), enquanto GEIG é só global

Diferença entre objectivos e métodos

- Objectivos: linguísticos, informáticos, etc.
- Métodos: qualitativos, quantitativos, estatísticos, simbólicos, sociológicos, filosóficos, introspectivos, empíricos, neurológicos, ... políticos, militares, religiosos...

Só para dar um cheirinho...

- Avaliar intuições em relação à língua
- Avaliar inferências legítimas

- Engenho na forma de estabelecer a tarefa
- Engenho no relacionamento das hipóteses
- Engenho na forma de analisar o resultado

Localização

- Qual a diferença entre
  - no canto da sala/ao canto da sala
  - por baixo da mesa/debaixo da mesa
- A comida está na mesa
- A tampa está na mesa
- O livro está em cima da mesa
- O sofá está à frente do armário
- A mesa está à frente do sofá

Web Track TREC 2003 (Craswell et al)

- destilação de um tópico:
  - que páginas são as mais importantes sobre aquele tópico?
  - RI+sumarização?
- encontrar uma página (não necessariamente de topo)

- CLIR: que modelo de utilizador?

HAREM: avaliação pormenorizada

- Quantos casos são fáceis?
  - basta procurar no almanaque
  - é sempre a mesma coisa
- Quantos casos exigem desambiguação?
  - entre categorias
  - entre serem ou não EMs
- Quantos casos são vagos?
- Quantos casos são criativos? (difíceis)